PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Assembly Theory is an approximation to algorithmic complexity based on LZ compression that does not explain selection or evolution

Felipe S. Abrahão[1,2,3], **Santiago Hernández-Orozco**[1], **Narsis A. Kiani**[4,5], **Jesper Tegnér**[6], **Hector Zenil**[1,5,7,8,9¤*]

1 Oxford Immune Algorithmics, Oxford University Innovation, Oxford, United Kingdom, **2** Center for Logic, Epistemology and the History of Science, University of Campinas (UNICAMP), Brazil, **3** DEXL, National Laboratory for Scientific Computing (LNCC), Brazil, **4** Department of Oncology-Pathology, Center for Molecular Medicine, Karolinska Institutet, Sweden, **5** Algorithmic Dynamics Lab, Center for Molecular Medicine, Karolinska Institutet, Sweden, **6** Living Systems Lab, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, **7** The Alan Turing Institute, British Library, London, United Kingdom, **8** School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom, **9** King's Institute for Artificial Intelligence, King's College London, United Kingdom

¤Current address: School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom
* hector.zenil@kcl.ac.uk

Citation: Abrahão FS, Hernández-Orozco S, Kiani NA, Tegnér J, Zenil H (2024) Assembly Theory is an approximation to algorithmic complexity based on LZ compression that does not explain selection or evolution. PLOS Complex Syst 1(1): e0000014. https://doi.org/10.1371/journal.pcsy.0000014

Editor: Hocine Cherifi, Université de Bourgogne: Universite de Bourgogne, FRANCE

Received: May 2, 2024

Accepted: August 23, 2024

Published: September 23, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000014

Data Availability Statement: There was no data being used in this research.

Funding: This work was supported by the São Paulo Research Foundation (FAPESP), grants

Abstract

We formally prove the equivalence between Assembly Theory (AT) and Shannon Entropy via a method based upon the principles of statistical compression that belongs to the LZ family of popular compression algorithms. Such popular lossless compression algorithms behind file formats such as ZIP and PNG have been shown to empirically reproduce the results that AT considers its cornerstone. The same results have also been reported before AT in successful application of other complexity measures in the areas covered by AT such as separating organic from non-organic molecules and in the context of the study of selection and evolution. We demonstrate that the assembly index is equivalent to the size of a minimal context-free grammar. The statistical compressibility of such a method is bounded by Shannon Entropy and other equivalent traditional LZ compression schemes, such as LZ77 and LZW. We also demonstrate that AT, and the algorithms supporting its pathway complexity, assembly index, and assembly number, define compression schemes and methods that are subsumed into algorithmic information theory. We conclude that the assembly index and the assembly number do not lead to an explanation or quantification of biases in generative (physical or biological) processes, including those brought about by (abiotic or biotic) selection and evolution, that could not have been arrived at using Shannon Entropy, or that have not been already reported before using classical information theory or algorithmic complexity.

2021/14501-8 (to FSA) and 2023/05593-1 (to FSA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Assembly Theory (AT) has recently been proposed in order to investigate the distinction between abiotic from biotic matter, while explaining and quantifying the presence of biosignatures, selection, and evolution. We previously have shown that AT cannot rule out false positives and that it has equal or worse performance in comparison to popular compression algorithms at counting exact copies in data without evidence that their compression mechanics are favoured over others. This article investigates these limitations and the many challenges of the theoretical foundations of AT. We demonstrate that AT's complexity measures (both for individual assembled objects and ensembles of objects) are subsumed into algorithmic information theory. The calculated assembly index for an object in AT is equivalent to the size of a compressing context-free grammar, and its calculation method is an LZ compression scheme that cannot perform better than Shannon Entropy in stochastic scenarios and cannot deal with non-stochastic (generative/causal) ones. Although AT may contribute with a graph-like pedagogical approach to LZ compression in application to molecular complexity, this article disproves hyperbolic claims raised by the authors of AT that introduce AT as a novel method, fundamentally different from other complexity indexes, or as a breakthrough. Instead, the principles behind AT are known elementary principles of complexity rehashed but introduced high logical inconsistency. AT lacks empirical evidence that it is different from or outperforms other complexity indexes in connection to selection, evolution or any of the applications in which the authors of AT have promoted it as capable of explaining physical and biological phenomena.

1 Introduction

Assembly Theory (AT) is a hypothesis that has recently garnered significant attention. Although its hyperbolic claims have not been taken seriously, the scientific media has amplified the many misleading arguments of their authors. Responding to how open-ended forms can emerge from matter without a blueprint, AT purports to explain and quantify the presence of biosignatures, selection, and evolution, and has recently even suggested to be able to explain time, matter, the expansion of the universe and even cosmic inflation [1]. The assembly index (or MA, a variation applied to quantify the assembly index on molecules) is the proposed complexity measure used to allegedly distinguish abiotic from biotic matter, the central claim of AT as advanced in [2, 3]. According to the assembly index, objects with a high assembly index "are very unlikely to form abiotically". This has been contested in [4], whose results "demonstrate that abiotic chemical processes have the potential to form crystal structures of great complexity", exceeding the assembly index threshold proposed by AT's authors. The existence of such abiotic objects would render AT's methods prone to false positives, corroborating the predictions in [5]. The assembly index (or MA) was shown to perform equally well, or worse in some cases, relative to popular compression algorithms (including those of a statistical nature) [5], some of which have been applied before in [6–8].

In order to investigate the fundamental limitations and theoretical underpinnings of such a complexity measure, and to explain the above results, this article shows that the non-linear (or "tree"-like) structure of the minimum rooted assembly (sub)spaces (from which assembly indexes are calculated) is one of the ways to define a (compression) scheme that encodes the assembling process itself back into a linear sequence of codewords (or phrases). We

demonstrate that the assembling process that results in the construction of an object according to AT is a *compression scheme* that belongs to the *LZ family* of compression methods, and therefore cannot perform better than *Shannon Entropy* in stochastic scenarios. All the detailed formalism and proofs are provided in the *supporting information* of the present article (S1 Appendix).

The LZ family includes many distinct compression methods, such as LZ77, LZ78, LZW, LZSS, and LZMA [9]. In the same manner as exemplified by the illustrative individual cases in later work in [10], one can trivially prove, for example, that the parsing according to the LZ78 differs from the parsing according to the LZ77, or that the one according to the LZMA differs from LZW, and so on. In accordance with the results earlier demonstrated in the present article, the examples presented in [10] further confirm that different parsings give rise to variations of compression schemes in the LZ family. The pointers in traditional LZ compression schemes, such as LZ77, LZ78 or LZW, refer to encoded tuples that have occurred before as recorded by the associated dictionary. Instead, for minimum rooted assembly subspaces, the pointers refer to other tuples that might not have occurred before in a particular assembling path, but are part of the minimum rooted assembly subspace, that is, part of a "tree"-like structure that is encoded in the dictionary, which is only built and queried during the encoding/decoding process.

The compressibility achieved by the *LZ scheme* that the *assembly index* calculation method is equivalent to depends on the length of the shortest assembling paths and the "simplicity" of the minimal assembly spaces. This notion of simplicity encompasses both how the assembly space (AS) structure differs from a single-thread (or linear) space and how many more distinct assembling paths can lead to the same object. Such an LZ scheme reduces the complexity (which the assembly index aims to quantify) of the assembling process of an object to the compressibility and computational resource efficiency of a compression method.

The generative/assembling process in an assembly space (AS) is strictly equivalent to a *Context-Free Grammar* (CFG), while the assembly index is equivalent to the size of such a compressing grammar. This implies that the assembly index can only be an approximation to the number of factors (or phrases) [11] in LZ schemes (e.g., LZW), which are statistical compression methods whose compression rates converge to that of Entropy. Thus, the assembly index calculation method is also a CFG-based compression method whose statistical compression rate is bounded by LZ schemes, such as LZ78 or LZW, and therefore by Entropy.

The theory and methods of AT describing a compression algorithm are, therefore, subsumed into Algorithmic Information Theory (AIT), which in turn has been applied in the same areas that AT has covered, from organic versus non-organic compound classification to bio- and technosignature detection and selection and evolution [6, 7, 12, 13]. This makes the assembly index and assembly number proposed by AT effectively approximations to algorithmic (Solomonoff-Kolmogorov-Chaitin) complexity [14–17].

Algorithmic complexity has been shown to be profoundly connected to causality [18–21], and not only in strings. It has also been applied to images [13], networks [22–25], vectors, matrices, and tensors in multiple dimensions [26–28]; and to chemical structures [6]. In contrast to the statistical compression schemes on which AT is based, and whose assembly index method is a particular case of a compression scheme, exploring other resource-bounded computable approximations to algorithmic complexity [20, 29] that consider aspects other than traditional statistical patterns such as identical repetitions, may have more discriminatory power. This is because they may tell apart cases with a high assembly index and an expectedly low copy number (or frequency of occurrence in the ensemble) from those with low LZ compressibility and high Entropy, seeming to indicate that an object is more causally disconnected, independent, or statistically random, while actually, it is strongly causally dependent due to

non-trivial rewriting rules not captured by statistical means [5, 30]. This is because AT is to statistical correlation, as measures of algorithmic complexity are to causation. Rarely the two are the same—correlation is not causation—and when they are, they are fully captured by other traditional statistical measures such as Shannon Entropy without the introduction of a methodological different framework that adds no more to the correlation problem beyond Shannon Entropy.

As proposed in [3], the *assembly number* is intended to measure the amount of selection and evolution necessary to produce the ensemble of (assembled) objects. That is, the assembly number aims to quantify the presence of constraints or biases in the underlying generative processes (e.g., those parts of the environment in which the objects were assembled) of the ensemble, processes that set the conditions for the appearance of the assembled objects. A higher assembly number—not to be conflated with the assembly index—would mean that more "selective forces" were in play as, e.g., environmental constraints or biases, in order to allow or generate a higher concentration of high-assembly-index elements. Otherwise, these high-assembly-index objects would not occur as often in the ensemble.

Consonant with the constraints and biases that the assembly number aims to quantify, though in fact, as we demonstrate, it constitutes a *compression method* subsumed into AIT, ensembles with higher assembly numbers are more compressible and would therefore diverge more markedly from those ensembles that are outcomes of (or constituted by) perfectly random processes. In turn, a more compressible ensemble implies that more constraints or biases played a role in generating more high-complexity objects more frequently than would have been the case in an environment with fewer constraints and biases (i.e., a more random or incompressible environment), thus increasing the frequency of occurrence of less compressible objects in this environment. Should one *assume* that the assembly number is indeed capable of measuring this feature or characteristic of the ensemble as a whole from the distribution of the assembled objects, then a higher assembly number would mean that more constraints or biases played a role in increasing the frequency of occurrence of high-assembly-index objects.

Conversely, under the same assumption, the presence of more biotic processes in an ensemble would imply a higher assembly number, which in turn imply that the ensemble is more compressible. This occurs, for example, in scenarios where there is a stronger presence of *top-down* (or downward) *causation* [21] behind the possibilities or paths that lead to the construction of the objects, while a less compressible ensemble would indicate a weaker presence (or absence) of top-down causation (see also Section 3).

We therefore conclude that AT cannot offer a different or better explanation of selection, evolution, or top-down causation than the connections already established [7, 18, 21], consistent with our previous position that a single, intrinsic scalar is unlikely to classify life or quantify selection in evolutionary processes independently of the environment and the perturbations it imposes on the objects (whether these are biotic or abiotic). Characterising the complexity of the phase transitions in complex systems has long been investigated within the scope of emergent and self-organising properties that either derive from or downwardly affect interactive relationships between local (micro-level or individual) systems [21, 31–33]. In order to understand the interactions between the system's parts and the environment's subsystems, it is necessary to devise sound theories that explain crucial aspects of complex systems in general, such as swarm intelligence [34], gene regulatory networks [35], self-organisation [31, 32, 36], and two-way causality effects in feedback loops between the system and the environment [21, 37, 38].

An intrinsic complexity measure, such as the one put forward by AT, would also face many challenges to explain phase transitions near the criticality, where systems are on the verge of collapsing ordered patterns into disorganised parts or stochastic randomness [36, 39]. This is

because the investigation of collective dynamics that accounts for other extrinsinc factors, e.g. those brought about by causal effects between multiple subsystems [20], is a challenge to such intrinsic complexity measures. These would miss the quantification of phenomena resulting from network dynamics that are crucial to understand (abiotic or biotic) complex systems [6, 22, 40, 41], whose emergent increase in the collective/global computation capabilities has been shown to be fundamentally connected to network topological properties [21, 42, 43].

As shown in the following Sections 2.1, 2.2, 2.3, 2.4, and 2.5, and demonstrated in S1 Appendix, the claim advanced by its authors that AT unifies life and biology with physics [3, 44] relies on a circular argument and on the use of a popular compression algorithm, with no empirical or logical support to establish deeper connections to selection and evolution than those already known (such as high modularity and self-assembly), already made (in connection to complexity) [7, 21, 22], or previously investigated using information- and graph-theoretic approaches to chemical and molecular complexity [6, 45–47].

2 Results

2.1 The assembly index is a compression algorithm of the LZ family

Despite the authors' assertion in the Assembly Theory (AT) paper that this theory and the assembly index are unrelated to algorithmic complexity, it is evident that AT is fundamentally encompassed within the realm of algorithmic complexity [5]. The assembly index, as proposed, seeks to gauge the complexity of an object based on the number of steps in its shortest copy-counting assembly pathway [3] via a procedure equivalent to LZ compression, which in turn is a computable approximation to algorithmic complexity, denoted by **K**.

At its core, as demonstrated in the S1 Appendix (Section A.3.3), the assembly index calculation method belongs to the *LZ family* of compression algorithms [48, 49], sharing the same key ideas that make LZ schemes converge to Shannon Entropy at the limit: the identification of repeated blocks, the usage of a dictionary containing repetitions and substitutions, and the ability to losslessly reconstruct the original object using its minimal LZ description. The assembly index calculation method (namely, 'LZAS'), as all of the compression methods in the *LZ family*, is defined by a scheme that resort to a (static or dynamic) dictionary containing tokens, indexes, pointers, or basic symbols from which the decoder or decompressor can retrieve the original raw data from the received encoded/compressed form according to the respective LZ scheme [9].

As also formally demonstrated in the S1 Appendix (Section A.3.3), the parity in number of steps for compression and decompression using the assembly index effectively reduces its definition in AT to the non-linearity of assembly spaces and the length of the assembling paths, reduction which in turn constitutes a loose upper bound of a resource-bounded approximation to **K** (see S1 Appendix, Sections A.2 and A.3.1).

In contrast, more robust approximations to **K**, capable of capturing blocks and other causal content within an object, have been proposed for purposes ranging from exploring cause-and-effect chains to quantifying object memory and characterising process content [29, 50]. These more advanced measures have found application in the same contexts and domains explored by AT, encompassing tasks like distinguishing organic from non-organic molecules [6], investigating potential connections to selection and evolution [7, 22], the detection of bio- and technosignatures [51, 52], and explorations into causality [18, 20]. Importantly, when applied to the data employed as evidence by AT, the more sophisticated measures consistently outperform the assembly index (see S1 Appendix, [5], and [53]). Thus, it becomes evident that AT and its assembly index represent a considerably constrained version of compression algorithms, and a loose upper bound of **K**.

This limitation is attributable to the authors' exclusive consideration of computer programs adhering to the form of 'Template Program A', as follows:

$$\textbf{Template Program A}: \begin{array}{l}\text{'while end-of-object, do } N \text{ times print(repetitions)}+\\ \text{print (all remaining objects not found in repetitions)'}\end{array}$$

In a published paper [54], the authors offered a proof of computability of their assembly index to distance themselves from algorithmic (Kolmogorov) complexity. Their algorithm, categorised as the above **Template Program A**, is trivially computable and requires no proof of computability, but all other resource-bounded approximations to **K** are also computable, including LZW that has been used for 60 years for similar purposes [16].

Any approximation to **K** that accounts for identical repetitions, including all known lossless statistical compression algorithms, can achieve equivalent or superior results, as demonstrated in S1 Appendix and [5]. This alignment with 'Template Program A' effectively highlights the association of AT with well-established principles of compression and coding theory, thereby refuting the initial claim of its authors to present a unique methodology different from other complexity measures, those which have been already applied to the same purposes of AT. Moreover, the authors' suggestion that their index may be generalised as a universally applicable algorithm for any object (including text) [55] further underscores the disconnect between AT—and its authors' drive to reinvent traditional algorithms such as text compression based on Shannon Entropy—and the current state of the art in the field of statistical and non-statistical compression beyond LZW [18, 30, 40].

Notice that although there are variations in the implementation of the assembly index (and the assembly number), of which the authors themselves have proposed significantly different versions to deal with the intrinsic intractability of their methods—see also Section 3 and S1 Appendix—, we demonstrate in S1 Appendix (Section A.3.3) that they are all qualitatively and quantitatively equivalent to the LZ algorithms, a family of compression schemes introduced in the 1970s, such as LZ77/LZ78. 'LZAS' (which is the LZ encoding-decoding scheme to which the assembly index calculation method is equivalent) translates the notion of the assembling paths playing a role in how much simpler the object can get, i.e., as we have demonstrated, how much more compressible the object is. We demonstrate that the notion of "simplicity" grasped by the assembly index encompasses both how much the assembly space structure differs from a single-thread (or linear) space and how many more distinct assembling paths can lead to the same object. If the search for the minimal assembly subspaces needs to cover a wider space of possible assembling paths in order to calculate the assembly index, then this process may turn out to be computationally expensive. To tackle this intractability, AT employed some approximation methods to the calculation of the assembly index, such as the split-branch version [54, 56], thus taking advantage of a narrower search by constraining the potential assembly subspaces. Our proofs in S1 Appendix (Section A.3.3) clarify the fact that any (whether more computationally efficient or not) approximation method, e.g. the split-branch version [54, 56], not only is an *LZ scheme* but also a variation of 'LZAS' that (in the best case scenario) can only improve on computational resources costs and/or compression rate: the more efficiently the approximation method gets closer to the (actual, but intractable in the general case) value of assembly index, the more the corresponding variation of (the actual) 'LZAS' improves on compression rate and/or the usage of computational resources to achieve such a compression rate.

In essence, the assembly index is fundamentally underpinned by the LZ encoding of the objects it measures. This reveals that the assembly index, and consequently AT, aligns more closely with the principles of traditional information theory (Shannon Entropy), and statistical

data compression than the authors are willing to acknowledge [54]. In fact, as demonstrated in S1 Appendix (Sections A.2 and A.3), both AT and the statistical compression methods that underpin it are subsumed into the algorithmic information theory.

AT's authors also assert that their index's ability to differentiate between organic and non-organic compounds validates its natural applicability. However, when compared with other statistical indexes, including various compression algorithms, these alternative methods often result in similar or superior performances [5]. This undermines the claims made in favour of AT.

2.2 A compression algorithm is a mechanical procedure that corresponds to a physical process

The authors of AT argue that traditional data compression algorithms are an overly abstract process unsuited for modelling the construction (or assembly) of objects [10, 55]. First, this view overlooks the fact (see Sections 2.1 and 3) that a minimal rooted assembly subspace is also an abstract compression scheme (in particular, in the LZ family) as much as any other compression scheme found in the literature, but one that in particular is intractable both in principle and in practice—so that in order to tackle such a limitation, AT has to employ approximation techniques. See also S1 Appendix (Section A.3.3). Secondly, this view overlooks the practical and mechanistic nature of compression algorithms, particularly those in the LZ family. Since their introduction in 1977, LZ algorithms have been effectively used in detection, identification, clustering, and classification across various fields, including biology, chemistry, and medicine [6, 57, 58], and to approximate algorithmic complexity **K** [16, 30]. Thirdly, traditional LZ compression schemes such as LZW or LZ77 are computationally efficient in principle and in practice.

The argument presented by the proponents of AT regarding the uncomputability of algorithmic complexity is deeply misguided. While it is true that **K** is semi-computable, computable (resource-efficient) algorithms like LZ77/LZ78/LZW have been: (i) widely used to approximate it; (ii) and applied in biology and chemistry to the same purposes of AT, challenging the assertion that AT represents a unique or superior approach [5, 6, 53]. Also notice that in a 1976 article [59], Lempel and Ziv defined an early version of their algorithm directly as a computable method to approximate algorithmic complexity, defining what is known as LZ complexity. This reveals that not only in its future applicability on estimating the complexity of objects but also in the intention to be a complexity measure, the intentions (or motivation) of AT [10] are in fact similar to those behind LZ compression algorithms (for more discussion, see also Section 3 and S1 Appendix). One distinction is that the latter was *not* initially intended to measure molecular complexity, but it was explicitly intended to be an applicable complexity measure to approximate algorithmic complexity; while the former was explicitly intended to measure molecular complexity, but it misses the fact that is also subsumed into, and inspired by, algorithmic complexity—to which AT's methods are approximations. Assuming that one does *not* recur to a wishful thinking fallacy [10], the merit of the applicability of any mathematical model or scientific theory should be judged according to its empirical significance in comparison to other approximation methods—in particular, all of them already suboptimal with respect to algorithmic complexity, AT included—in the same field; and should *not* be judged by the intentions of the scientists that proposed a theory. (See Section 3 for more discussion on *future research*).

The emphasis on the purported requirement of a Turing machine in the context of algorithmic complexity is a misdirected concern. 'Turing machine' is synonymous with 'algorithm' and an 'algorithm' is a synonym for a 'Turing machine'. Any formal mathematical theory and

method that is algorithmic in nature, including all the (computable) methods in AT, are algorithms, and are therefore technically Turing machines or programs that can run on a (universal) machine. In addition, this holds regardless of the objects themselves (or the underlying processes that generate or govern them) being actual Turing machines, or not.

Further evincing such a misguided argument, it is well-known that computation is a process that not necessarily is performed by a Turing machine. For example, context-free grammars (CFGs) and pushdown automata perform computation, while being fundamentally different (by construction and in theory) from a Turing machine: their mathematical definition is distinct from that of a Turing machine; and they are proven to belong to distinct computational classes—particularly, a proper subclass below the hierarchy of recursive functions. Now, consider the *premise*: the assembling process of objects as formalized by AT (in the form of assembly spaces) is a mathematical scheme or model that does *not* model any type of computation process (whether physical or abstract). As we demonstrate in S1 Appendix (Section A.3.2), assembly (sub)spaces are formally equivalent to CFGs, and additionally the assembly index itself is equivalent to the size of the corresponding minimal compression CFG. Then, according to such a premise, CFGs also would *not* model any type of computation process. However, one already knows that the latter is false, therefore demonstrating a fundamental *contradiction* in such a premise.

Furthermore, in case AT presents empirical significance in distinguishing the abiotic-to-biotic transition, so it will CFGs. Notice that the fact that AT's methods are contained in a particular subrecursive class was already proved in [5], and in S1 Appendix (Section A.3.3) we also demonstrate that the *compression rate* defined by 'LZAS' (i.e., the assembly index calculation method) is indeed dependent on the length of the assembling paths in a minimal rooted assembly subspace. As expected from a member of the LZ family, such as the assembly index calculation method, the compressibility achieved by traditional members of the LZ family, such as LZW, similarly depends on the number of phrases or factors in the parsing of an object [60].

Such statements from the authors of AT [55] indicate a fundamental misunderstanding of equivalence classes closed under reductions (in this case, computability classes), which are basic and pivotal concepts in computer science, complexity theory, and mathematics in general. As a trivial exercise, and evidence in support of this point, is that none of the formal proofs in the S1 Appendix make any mention of any Turing machine. In the same manner, all Turing machines referenced in [5] can be, for example, replaced by any sufficiently expressive programming language running on an arbitrary (abstract or physical) computer.

On the contrary to the claims of AT's authors [2, 10, 55], one of the distinctive features of algorithmic information theory (AIT), algorithmic complexity being one of its indexes, is that its importance and pervasiveness in mathematics, theoretical computer science, and complexity science are in fact owed to the invariance of its results with respect to the choice of the language, formal mathematical theory, and the computation model, whether abstract or physically implemented.

In practice, computable approximations like LZ77/LZ78—which AT mimics without attribution—do not require a Turing machine, and have been widely and successfully used in clustering and categorisation [16, 57], including in the successful separation of chemical compounds into organic and non-organic categories [6], and in the reconstruction of evolutionary phylogenetic trees [57]. To say that a physical (computable/recursive) process like those affected by AT's methods is not a compression algorithm because it does not resemble or correspond to the functioning of a Turing machine is as naively wrong and misplaced as saying a program written in Python cannot model the movement of a pendulum because nature does not run Python. A Turing machine is an abstraction and a synonym for 'algorithm.'

Everything is an algorithm in Assembly Theory, and therefore it is governed by the same principles of computer science and information theory.

This type of argument also reflects a misunderstanding of the purposes of Turing machines and of the foundations of computer science. As explained in a quote often attributed to Edsger W. Dijkstra,

"Computer science is no more about computers than astronomy is about telescopes."

At its inception, Turing machines were defined within computer science as an abstraction of the concept of an algorithm, years before what we now know as computers were built.

2.3 Conflating object assembly process and directionality of causation

The authors of Assembly Theory (AT) assume and present the sequential nature of their algorithm as an advantage [55]. AT claims to be able to extract causal knowledge by measuring the degree of causality in the form of non-unique chains of cause and effect across the assembly pathways [55]. The central assumption is that each step of the basic **Template Program A** would constitute a cause-and-effect chain corresponding to how an object may have been physically assembled from an assembly path, which may not be unique. This is no different from, and is indeed a restricted version of, a (deterministic or non-deterministic) pushdown automaton capable of instantiating a grammar compressor (which the authors rename a minimal rooted 'assembly subspace') and works exactly like an LZ algorithm. See S1 Appendix (Sections A.3.2 and A.3.3).

The assumption of sequential assembly from instantaneous object snapshots is not based on any physical (biological, chemical, or other) evidence. In contrast, there is overwhelming evidence that this is not the case. For example, in the case of a genome sequence, all the regions of the genome sequence are exposed to selective forces simultaneously. Transcription factors, or genes that regulate other genes, do not get assembled or interact only sequentially. In chemistry, reactions do not happen on one side of a molecule first and then propagate to the other but happen in parallel. Reactions in time follow and are represented sequentially. However, throughout AT's arguments, causal directionality in time is conflated with how an object may have been assembled from its instantaneous configuration.

This assumption of sequential assembly of an object based on causal direction is manifestly incorrect in regards to how an object subject to selection assembles. In contrast, as further discussed in Section 3, only by finding the generative mechanism of the object—such as the underlying set of (computable) mechanisms (of which the assembly index only takes into account a particular case)—and thus explaining the object in a non-trivial fashion, can one reproduce both the causal direction from the sequence of connected steps (see Fig 1B) and how the object itself may have been assembled (which is not and cannot be in a sequential fashion). AT and its index cannot characterise this, for example, because the class of problems or functions recognised by pushdown automata (or generated by context-free grammars) is a proper subset of the class of recursive problems.

The authors of AT have mistakenly claimed that algorithmic complexity and Turing machines are not, or cannot be, related to causality. This is incorrect [18, 20]. They were introduced as causal artifacts for studying mechanistic means of performing logical operations.

As previously presented in Section 2.2, we have demonstrated in S1 Appendix (Section A.3.2) that the assembly index requires a finite automaton to instantiate a context-free grammar (CFG), a version of a specific-purpose machine, just as any other resource-bounded approximation to **K** would need to be instantiated (e.g. calculated, even by hand). In this case,

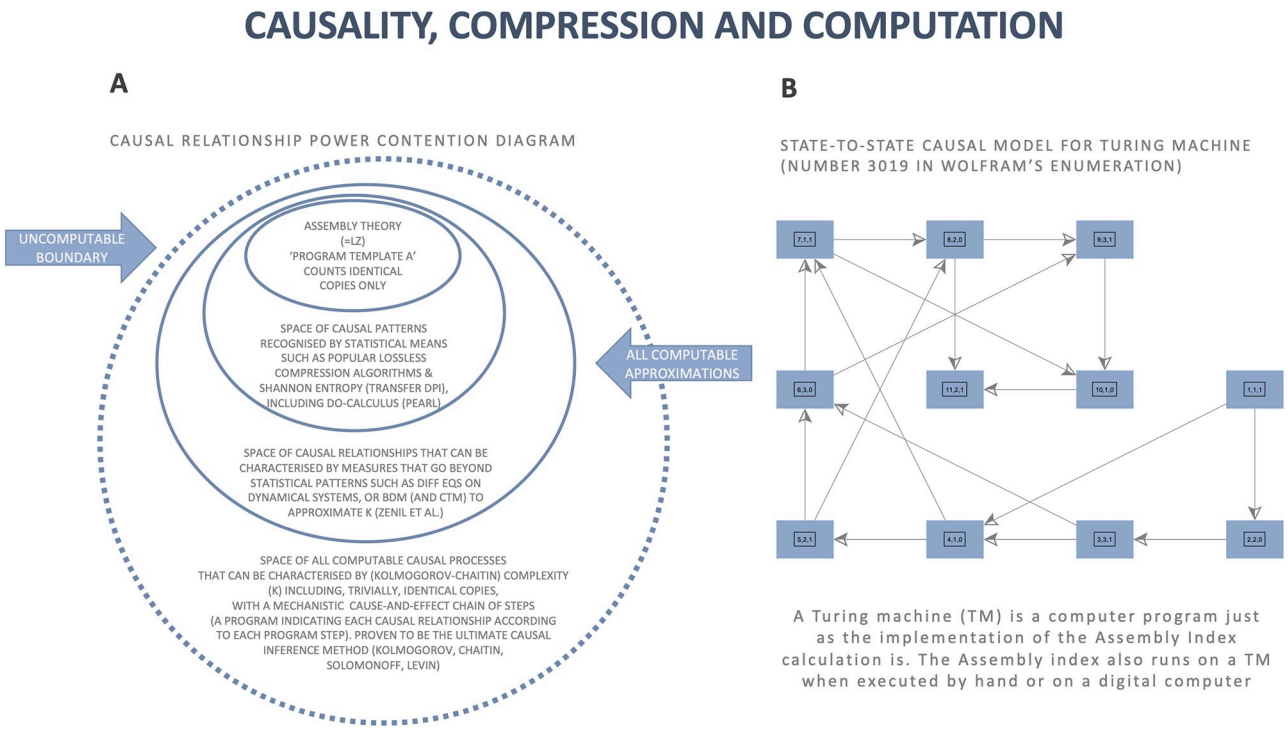

Fig 1. A: The authors of AT have suggested that algorithmic complexity (**K**) would be proven to be contained in AT [55]. This Venn diagram shows how AT is connected to and subsumed within algorithmic complexity and within the group of statistical compression as proven in this paper (see S1 Appendix) by a simple template argument representing the very restricted type of complexity that is able to capture. B: Causal transition graph of a Turing machine with number 3019 (in Wolfram's enumeration scheme [36]) with an empty initial condition found by using a computable method (e.g. CTM [62]) to explain how the block-patterned string 111000111000 was assembled step-by-step based on the principles of algorithmic complexity describing the state, memory, and output of the process as a fully causal mechanistic explanation. A Turing machine is simply a procedural algorithm and any algorithm can be represented by a Turing machine. By definition, this is a mechanistic process as originally intended by Alan Turing himself, and as physical as anything else (first computers were human), not an 'abstract' or 'unrealisable' process as the authors of AT have suggested [54] misunderstanding a basic concept.

https://doi.org/10.1371/journal.pcsy.0000014.g001

the size of their formula or the size of the implementation of their computer program is the size of the special-purpose finite automaton, which is common and of fixed size for their calculations (which allows them to be discounted from the final length, just as it is from **K** or resource-bounded approximations).

That the calculation of the assembly index from the minimum rooted assembly spaces can be reduced to a grammar compression scheme demonstrates not only that the assembly process' complexity is dependent on such contingencies as which assembly spaces or which distinct paths were taken in past historical stages of building the object, but also that this recursivity and dependence on past trajectories are subsumed into the concept of compressibility. However, unlike algorithmic approaches, assembly theory does not incorporate any of the elements of causality such as perturbation and counterfactual analysis, as has been done in the context of algorithmic complexity [18, 21, 22], a generalisation that when compared to AT reflects the concealed simplistic nature of AT.

2.4 Lack of control experiments and absence of supporting empirical data beyond current domain knowledge

A critical examination of the AT methodology reveals significant shortcomings. Firstly, proponents of AT failed to conduct basic control experiments, a foundational aspect of introducing

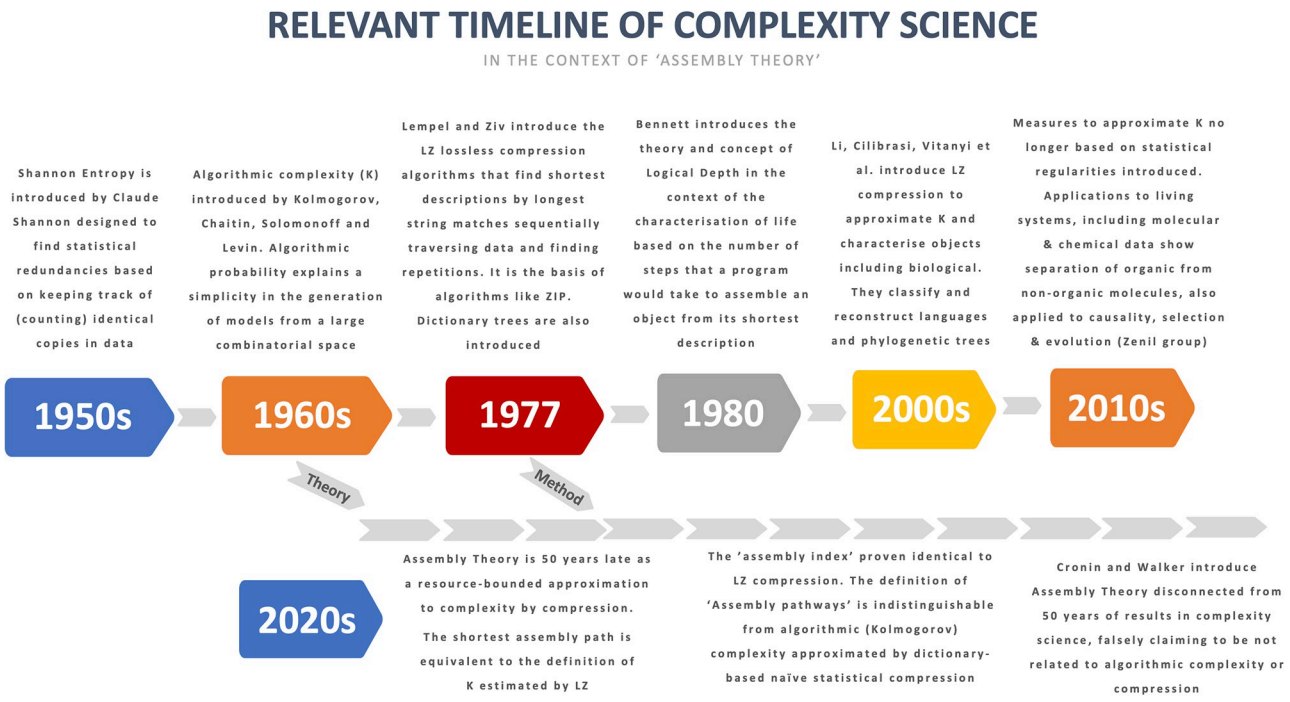

Fig 2. A timeline of results in complexity science relevant to the claims and results of AT, which renames several concepts, e.g. dictionary trees as 'assembly (sub)spaces'; relies heavily on algorithmic probability in its reduction of combinatorial space arguments, without attribution; and, as demonstrated, the assembly index is an LZ compression scheme (proofs provided in the S1 Appendix).

https://doi.org/10.1371/journal.pcsy.0000014.g002

a new scientific metric. Benchmarking against established indices, particularly in coding and compression algorithms, is crucial to validating any new metric in the domain. Previous work on AT has never included meaningful experimental comparisons of the assembly index with other existing measures on false grounds that their measure is completely different [10, 54] (see Figs 1 and 2). Yet, we have shown that other algorithms, such as RLE, Huffman coding (the first dictionary-based universal compression algorithm), and other compression algorithms based on dictionary-based methods and Shannon Entropy produced equivalent or superior results compared to the results published by the authors of AT [5].

AT also introduces a cutoff value in its index, purported to offer a unique perspective on molecular complexity, distinct from those based on algorithmic complexity. This value indicates when an object is more likely to be organic, alive, or a product of a living system. However, various indices and compression algorithms tested have yielded equivalent cutoff values (see [5]). In fact, such an Assembly Index threshold replicates previous results on molecular separation that employ algorithmic complexity. Thus, this overlap in results challenges the notion that AT provides a unique tool for distinguishing between organic and non-organic matter. For instance, similar cutoff values have been derived in other studies, such as those mentioned in [6], effectively separating organic from non-organic molecular compounds. While the authors of AT ignored decades of research in chemical complexity based on graph theory and the principles of Entropy [45–47, 61], the depth of the purported experimental validation of AT has been notably limited, especially when contrasted with more comprehensive studies. For instance, a more exhaustive and systematic approach, with proper control experiments involving over 15,000 chemical compounds [6] and employing algorithms from the LZ

family (or Entropy) and others of greater sophistication (BDM), demonstrated the ability to separate organic from non-organic molecular compounds.

Turing's motivation was to explore what could be automated in the causal mechanisation of operations in Fig 1B, for example, stepwise by hand using paper and pencil. The shortest among all the computer programs of this type is an upper bound approximation to **K**. In other words, **K** cannot be longer than the length of this diagram. The concept of pathway complexity and the algorithm for copy number instantiated by the assembly index formally represent a restricted version of a Turing machine or finite automaton. As also discussed in Section 2.2, it is a mistake to think of a Turing machine as a physical object or an object with particular properties or features (such as a head or a tape). All the algorithms in assembly theory are Turing machines.

2.5 A circular argument cannot unify physics and biology

Central to Assembly Theory (AT) and the public claims made through the authors' university press releases [44] is the connection that they make to selection and evolution, maintaining that AT unifies physics and biology and explains and quantifies selection and evolution, both biotic and abiotic [3, 10]. To demonstrate this connection, assuming selectivity in the combination of linear strings (or, equivalently, objects) (P), the authors compare two schemes for combining linear strings Q, one random versus a non-random selection. This experiment yields observations of differences between the two (R), and the authors conclude that selectivity (S) exists in molecular assembly, and therefore that AT can explain it.

Observing that a random selection differs from a non-random selection of strings, the authors use this as evidence for "the presence of selectivity in the combination process between the polymers existing in the assembly pool".

Yet, assuming selectivity in the combination of strings and proceeding to say that a selection algorithm is expected to differ from a random one makes for a circular argument. Furthermore, the conclusion (S) reaffirms the initial assumption (P), resulting in a circular argument. In a formal propositional chain, it can be represented as $P \Rightarrow Q \Rightarrow R \Rightarrow S \Rightarrow P$, where the conclusion merely restates the initial assumption, lacking any validation or verification, empirical or logical, beyond a self-evident tautology.

Whether this sequential reasoning is relevant to how molecules are actually assembled is also unclear, given the overwhelming evidence that objects are not constructed sequentially and that object complexity in living systems is clearly not driven by identical copies only [63]. See also Section 3.

Nevertheless, the question of random *versus* non-random selection and evolution in the context of approximations to algorithmic complexity, including copy counting, was experimentally tested, empirically supported, and reported before in [7] following proper basic principles such as a literature search and review, control experiments (comparison to other measures), and validation against existing knowledge in genetics and cell biology.

Such circular arguments, lacking a foundation in chemistry, biology, or empirical evidence, lead the authors to propose the concept of "assembly time". The suggestion is a time-scale separation between molecule production (assembly time) and discovery time, aiming to unify physics and biology. Time, however, has always been fundamental in evolutionary theory, and claims about an object's historical contingencies are the foundations of evolutionary theory. Thus, AT revisits existing complexity science concepts described in the timeline provided in Fig 2 but without acknowledgment or attribution.

3 Discussion

The most popular examples used by the authors of Assembly Theory (AT) in their papers as illustrations of the way their algorithms work, such as ABRACADABRA and BANANA, are traditionally used to teach LZ77, LZ78, or LZW compression in Computer Science courses at university level. Dictionary trees have been used for pedagogic purposes in computer science for decades. In addition, beyond a notation '$c(x)$' for the formalisation of the assembly indexes (and assembly spaces) [54] that resembles that of the number of phrases (or factors) in the parsings of LZ78 in the proof presented in [60], they also share underlying key ideas in the relationship between these parsings and the probability of the phrases into which the objects are decomposed. This is because one of the central motivations for the assembly index as a complexity measure is that the probability of assembling paths should decrease as the respective assembly index (or, in the case of molecules, MA) of the object increases [2], unless there are environmental constraints or an extrinsic agent responsible for increasing the frequency of occurrence of high-assembly-index objects. Similarly, this appears as the key idea in the proof that LZ78 achieves optimal compression [60] for stationary and ergodic stochastic processes: sequences with a larger number of distinct phrases to which the pointers necessarily recur less often would have lower probability, and therefore they are less compressible; while sequences with fewer distinct phrases corresponds to a higher probability, and therefore they can be compressed more efficiently.

At first glance, the above aspects suggest some similarity between AT and compression algorithms, particularly those of a statistical nature based on recursion to previous states, copy counting, and the re-usage of patterns and repetitions. Indeed, such an interdisciplinary approach between physics, chemistry, and computer science is very beneficial and fruitful for science in general, specially in the field of complex systems science. However, in contrast to AT's authors claims, we have shown in the present article that the calculation of the assembly index—by finding a minimal assembly (sub)space (AS) rooted in the basis objects—in fact is a *compression scheme* belonging to the *LZ family* of compression algorithms, rather than merely being similar. The present study also theoretically establishes and extends the limitations previously investigated in [5].

We have formally proved that the minimum rooted AS from which the assembly index is calculated is a *compression scheme* (namely, 'LZAS') that belongs to the LZ family of compression algorithms.

This result also mathematically proves how the compression rates achieved by the assembly index calculation method depend on shortest assembly paths, assembly indexes, and assembly spaces. Additionally, it demonstrates that the assembly index calculation method is a dictionary-based compression method whose encoded dictionary enables one to generate/assemble the objects, for example via a *context-free grammar* (CFG). In fact, we also prove that the assembly index itself is equivalent to the size of a compressing CFG. Diving into the details of AT's algorithms and methods, we have revealed that they are equivalent to LZ encoding, and therefore to (Shannon) *Entropy*-based methods. All of the results in this article are fully detailed in the S1 Appendix.

Our results prove that an object with a low assembly index has high LZ compressibility (i.e., it is more compressible according to the LZ scheme), and therefore would necessarily display low Entropy when generated by i.i.d. stochastic processes (or low Entropy rate in ergodic stationary processes in general). In the opposite direction, an object with a high assembly index will have low LZ compressibility, and therefore high Entropy. The notion of how complex an assembling process needs to be in order to construct an object is equivalent to how much less compressible (by the LZ scheme to which the assembly index calculation method is equivalent)

and more computationally demanding (according to this encoding-decoding LZ scheme) this process is. This interdependence between compressibility and computational efficiency characterises the notion of *complexity* that the assembly index intends to measure, but that in fact pertains, as we have demonstrated, to an LZ compression scheme.

Thus, the methods based on AT, Shannon Entropy, and LZ compression algorithms are indistinguishable from each other with regard to the quantification of complexity of the assembling process, and the claim that these other approaches are incapable of dealing with AT's type of data (capturing 'structure' or anything that could supposedly not be characterised by Shannon Entropy or LZW) is inaccurate.

Kempes et al. [10] argue in later work that the assembly index is distinct from LZW and Huffman encoding by presenting counterexamples in which their parsings differ. These examples highlight parsing differences already subsumed into the general-case proofs in the S1 Appendix. The LZ family tree of compression schemes includes many distinct schemes, including LZ77, LZ78, LZW, LZSS, and LZMA [9], and as demonstrated in the present article it also includes the assembly index calculation method (to which we refer as 'LZAS'). These LZ schemes may differ not only in the parsing (or decomposition into factors or phrases) of the object, but also in performance. All of these compression methods are defined by schemes that resort to a (static or dynamic) dictionary containing tokens, indexes, pointers, or basic symbols from which the decoder or decompressor can retrieve the original raw data from the received encoded/compressed form according to the respective LZ scheme [9, 60]. Thus, raising an argument such as the one that the assembly index calculation method is not LZ compression because its parsings differ from those of LZW is as pertinent as arguing that LZW is not LZ compression because its parsings differ from those of LZ77.

Computing the exact value of the assembly index (i.e., the size of the minimum rooted assembly space) is known to be intractable in the general case so that one needs to employ approximation methods, such as the split-branch version [54, 56], to speed up the process by constraining the search over the possible assembling spaces or paths. Because the assembly index calculation method is an LZ compression scheme, our results also demonstrate that such an intractability not only arises in potential applications of AT to coding and compression, but also that it is the very intractability which *in fact is intrinsic* to AT, occurring whenever one tries to calculate the assembly index for an assembly space—a process for which AT had to employ approximation techniques, such as the split-branch assembly space.

Different computationally efficient implementations of these approximation methods will produce varying results that eventually diverge (i.e., they are *suboptimal with respect to* what AT formalised and proposed as the ideally true value) from the optimal/ideal assembly indexes in the general case (which remains computationally intractable). As we have demonstrated, each of these more efficient approximation methods defines a variation of the (computationally inefficient) LZAS scheme, variation which is in turn equivalent to a more tractable compression scheme in the LZ family. Thus, the more computationally efficient an empirical implementation of the assembly index approximation method is, the more efficient the LZ compression scheme that such a method defines, and hence is equivalent to.

Since previous work in [5] (and also further investigated in [53]) has shown that the assembly index method performs equally well, or worse in some cases, an extensive evaluation of the performance of their algorithms in contrast to other established methods in the literature is necessary in *future research*, should the proponents of the assembly theory aim to argue that their specific implementations outperform established, computationally efficient compression algorithms like LZW for their specific purposes.

We have also demonstrated that AT is subsumed into the theory and methods of *Algorithmic Information Theory* (AIT) as illustrated in Fig 1A and mathematically proven in the S1

Appendix. In addition, both theoretical concepts and empirical results in [2, 3, 54, 64] have been reported previously in earlier work in relation to chemical processes in [6], and to biology in [6, 7, 12, 13] but with comparison to other measures, including methods that go beyond traditional statistical compression and are connected to the concept of causal discovery [20, 22].

In cases which AT displays discriminatory power [2, 64], it is because of its connection to Shannon Entropy and statistical compression in general. Our results first demonstrate that in the case of pure *stochastic processes*, the assembly number (not to be conflated with the assembly index) is either a *suboptimal* or *an optimal* compression method *with respect to* what the noiseless source coding theorem establishes, therefore adding *no* advantage in comparison to (Shannon) entropy-based methods. Secondly, in the *general case*—including the pure stochastic one but also when the degree of stochasticity or determinism of the generative processes of the ensembles is unknown—, they demonstrate that the *assembly number* is a compression method, and thus an *approximation* to algorithmic complexity (or, equivalently, to algorithmic probability), but a *suboptimal* measure *with respect to* the more general methods from algorithmic information theory.

AT may pedagogically contribute introducing a graph-like representational approach to approximating LZ compression in the specific context of molecular complexity potentially making it more accessible to a broader and less technically-minded community. However, as proven here, it is fundamentally not different either methodologically or fundamentally from Shannon Entropy and lossless compression and is inaccurate and a bad practice to ignore, omit or imply the absence of previous work (e.g. [6–8, 57]). We believe that it is wrong to belittle and devalue the theories and ideas (Shannon Entropy, compression, and algorithmic complexity) which AT's indexes and measures are based on, and make disproportionate public claims related to the contributions of AT (with the media and the authors themselves going so far as to call it a theory that 'unifies biology and physics' and a 'theory of everything' and so on, e.g. [1]).

Both the assembly number and the assembly index are currently defined in such a way that *prevents AT to quantify*, *even in principle*, the downward effects of environmental influence and constraints on the assembling process of the objects, thus preventing it to quantify the presence of *top-down causation*. For example, these effects have been shown before to have a fundamental relationship with the notion of complexity [21, 32, 65, 66]. This is because the assembly index (not to be confused with the assembly number) of an object in its current mathematical formulation is an *intrinsic measure* not sensible to increases or decreases in the "complexity" of the ensemble as a whole (or, as AT purports, in the skewness toward high-"complexity" objects in the ensemble). The assembly number in its current state can only quantify the *global* variations of "complexity" resulting from individual *local* variations of the objects' "complexities" (i.e., as proposed by AT, their assembly indexes), or variations in the distribution of these individual values—therefore, in principle, only bottom-up causation. However, it cannot quantify the other way around, that is, quantify local variations of "complexity" resulting from global variations of "complexity".

In order to take into account top-down causation, the assembly index of an object itself would need to also depend on the multi-agent interaction dynamics (or organisation) of extrinsic factors, e.g. environmental catalytic conditions that may play a role in adding or reinforcing biases toward the formation of higher-assembly-index objects through less likely assembly pathways. Otherwise, the assembly index may classify a low-complexity molecule as being constructed by a more complex extrinsic agent, which is in fact of a much simpler nature (e.g. a naturally occurring phenomenon or the interaction of many abiotic subsystems in the environment).

That is, in case sufficiently complex environmental catalytic conditions play the role of this extrinsic factor (which increases the bias toward the construction of a more complex molecule), such a level of complexity would be completely missed by the capabilities of a simplistic measure such as the assembly index, thereby rendering it *prone to* false positives [4, 5].

This kind of influence (whether top-down or bottom-up) in the interplay between global and local scales is e.g. encompassed by complexity measures based on *algorithmic probability* (or, equivalently, algorithmic complexity) [18, 21, 22, 67]: when one approximates the algorithmic probability of either an individual or a collective (e.g., an object or a set), a greater quantity of possible underlying generative processes of different nature (whether intrinsic or extrinsic) are taken into account with the purpose of estimating the most probable candidate generative model.

Thus, a pervasive comparison in *future research* between assembly number and algorithmic-probability-based measures for group-controlled ensembles sampled from distinct environments may help quantify the presence of AT's false positives. As the assembly number (or, as we have demonstrated in this article, the compressibility) of an abiotic ensemble sufficiently increases for certain environmental conditions, this relative increment of *global* "complexity" may be sufficient to in turn facilitate the formation of *abiotic* molecules with relatively higher assembly index. By empirically measuring the presence of such an effect on the assemblage of individual objects, *future research* is necessary to quantify the presence of top-down causation —a problem also raised in [68]—, and therefore it will help corroborate or falsify the AT's formalism and assumption of a bottom-up-only intrinsic complexity measure being able to distinguish the transition from non-life to life.

We agree with AT's authors on the importance of understanding the origin and mechanisms of the emergence of an open-ended generation of novelty in complex systems [21, 69]. However, these efforts by the authors are undermined by hyperbolic claims, fallacious arguments, and lack of attribution and comparison with previous results in the literature.

We argue that the results and claims made by AT, in fact, highlight a working hypothesis that information and computation underpin the concepts necessary to explain the processes and building blocks of physical and living systems, concepts which the methods and frameworks of AT have been heavily inspired by [70, 71].

Supporting information

S1 Appendix. Supporting information of the article. Additional PDF file provided along with the article that is the supplementary information containing all the detailed formalism, results, and proofs.
(PDF)

Author Contributions

Conceptualization: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

Formal analysis: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

Investigation: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

Methodology: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

Project administration: Hector Zenil.

Supervision: Hector Zenil.

Writing – original draft: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

Writing – review & editing: Felipe S. Abrahão, Santiago Hernández-Orozco, Narsis A. Kiani, Jesper Tegnér, Hector Zenil.

References

1. Templeton. A bold new theory on why the universe keeps expanding | Lee Cronin); 2024. YouTube video. Available from: https://www.youtube.com/watch?v=cYliayfoSDk.
2. 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). https://doi.org/10.1038/s41467-021-23258-x PMID: 34031398
3. Sharma A, Czégel D, Lachmann M, et al. Assembly theory explains and quantifies selection and evolution. Nature. 2023; 622(1):321–328. https://doi.org/10.1038/s41586-023-06600-9 PMID: 37794189
4. Hazen RM, Burns PC, Cleaves HJ, Downs RT, Krivovichev SV, Wong ML. Molecular assembly indices of mineral heteropolyanions: some abiotic molecules are as complex as large biomolecules. Journal of The Royal Society Interface. 2024; 21(211):20230632. https://doi.org/10.1098/rsif.2023.0632 PMID: 38378136
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. https://doi.org/10.1038/s41540-024-00403-y PMID: 39112510
6. Zenil H, Kiani NA, Shang Mm, Tegnér J. Algorithmic Complexity and Reprogrammability of Chemical Structure Networks. Parallel Processing Letters. 2018; 28(01). https://doi.org/10.1142/S0129626418500056
7. Hernández-Orozco S, Kiani NA, Zenil H. Algorithmically probable mutations reproduce aspects of evolution, such as convergence rate, genetic memory and modularity. Royal Society Open Science. 2018; 5(8):180399. https://doi.org/10.1098/rsos.180399 PMID: 30225028
8. Zenil H, Minary P. Training-free measures based on algorithmic probability identify high nucleosome occupancy in DNA sequences. Nucleic Acids Research. 2019; 47(20):e129–e129. https://doi.org/10.1093/nar/gkz750 PMID: 31511887
9. Salomon D, Motta G. Handbook of Data Compression. London: Springer London; 2010.
10. Kempes C, Walker SI, Lachmann M, Cronin L. Assembly Theory and Its Relationship with Computational Complexity. arXiv Preprints. 2024;(arXiv:2406.12176).
11. Rytter W. Grammar Compression, LZ-Encodings, and String Algorithms with Implicit Input. In: Díaz J, Karhumäki J, Lepistö A, Sannella D, editors. Automata, Languages and Programming. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 15–27.
12. Zenil H. Turing Patterns with Turing Machines: Emergence and Low-Level Structure Formation. Natural Computing. 2013; 12(2):291–303. https://doi.org/10.1007/s11047-013-9363-z
13. Zenil H, Delahaye JP, Gaucherel C. Image Characterization and Classification by Physical Complexity. Complexity. 2012; 17(3):26–42. https://doi.org/10.1002/cplx.20388
14. Chaitin G. Algorithmic Information Theory. 3rd ed. Cambridge University Press; 2004.
15. Calude CS. Information and Randomness: An algorithmic perspective. 2nd ed. Springer-Verlag; 2002.
16. Li M, Vitányi P. An Introduction to Kolmogorov Complexity and Its Applications. 4th ed. Texts in Computer Science. Cham: Springer; 2019.
17. Downey RG, Hirschfeldt DR. Algorithmic Randomness and Complexity. New York, NY: Springer New York; 2010.
18. Zenil H, Kiani NA, Tegnér J. Algorithmic Information Dynamics: A Computational Approach to Causality with Applications to Living Systems. 1st ed. Cambridge University Press; 2023.
19. Zenil H, Kiani NA, Abrahão FS, Tegnér JN. Algorithmic Information Dynamics. Scholarpedia. 2020; 15(7):53143. https://doi.org/10.4249/scholarpedia.53143
20. Zenil H, Kiani NA, Zea AA, Tegnér J. Causal deconvolution by algorithmic generative models. Nature Machine Intelligence. 2019; 1(1):58–66. https://doi.org/10.1038/s42256-018-0005-0

21. Abrahão FS, Zenil H. Emergence and algorithmic information dynamics of systems and observers. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2022; 380 (2227). PMID: 35599568
22. Zenil H, Kiani NA, Marabita F, Deng Y, Elias S, Schmidt A, et al. An algorithmic information calculus for causal discovery and reprogramming systems. iScience. 2019; 19:1160–1172. https://doi.org/10.1016/j.isci.2019.07.043 PMID: 31541920
23. Zenil H, Kiani NA, Tegnér J. Quantifying loss of information in network-based dimensionality reduction techniques. Journal of Complex Networks. 2015; 4(3):342–362. https://doi.org/10.1093/comnet/cnv025
24. Kiani NA, Zenil H, Olczak J, Tegnér J. Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. Seminars in Cell & Developmental Biology. 2016; 51:44–52. https://doi.org/10.1016/j.semcdb.2016.01.012 PMID: 26851626
25. Zenil H, Kiani NA, Adams A, Abrahão FS, Rueda-Toicen A, Zea AA, et al. Minimal Algorithmic Information Loss Methods for Dimension Reduction, Feature Selection and Network Sparsification. arXiv Preprints. 2023;(arXiv:1802.05843).
26. Zenil H, Soler-Toscano F, Jean-Paul D, Gauvrit N. Two-dimensional Kolmogorov complexity and an empirical validation of the Coding theorem method by compressibility. PeerJ Computer Science. 2015; 1:e23. https://doi.org/10.7717/peerj-cs.23
27. Abrahão FS, Wehmuth K, Zenil H, Ziviani A. Algorithmic Information Distortions in Node-Aligned and Node-Unaligned Multidimensional Networks. Entropy. 2021; 23(7). https://doi.org/10.3390/e23070835 PMID: 34210065
28. Abrahão FS, Wehmuth K, Zenil H, Ziviani A. An Algorithmic Information Distortion in Multidimensional Networks. In: Benito RM, Cherifi C, Cherifi H, Moro E, Rocha LM, Sales-Pardo M, editors. Complex Networks & Their Applications IX. vol. 944 of Studies in Computational Intelligence. Cham: Springer; 2021. p. 520–531.
29. 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. 2018; 20(8):605. https://doi.org/10.3390/e20080605 PMID: 33265694
30. Zenil H. A review of methods for estimating algorithmic complexity: options, challenges, and new directions. Entropy. 2020; 22(6):612. https://doi.org/10.3390/e22060612 PMID: 33286384
31. Jantsch E. The Self-organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution. Systems Science and World Order Library. Oxford: Pergamon press; 1984.
32. Prokopenko M, Boschetti F, Ryan AJ. An information-theoretic primer on complexity, self-organization, and emergence. Complexity. 2009; 15(1):11–28. https://doi.org/10.1002/cplx.20249
33. Zenil H, Gershenson C, Marshall J, Rosenblueth D. Life as Thermodynamic Evidence of Algorithmic Structure in Natural Environments. Entropy. 2012; 14(11):2173–2191. https://doi.org/10.3390/e14112173
34. Bonabeau E, Dorigo M, Theraulaz G. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press; 1999.
35. Kauffman S. Understanding Genetic Regulatory Networks. International Journal of Astrobiology. 2003; 2(2):131–139. https://doi.org/10.1017/S147355040300154X
36. Wolfram S. A New Kind of Science. Champaign, IL: Wolfram Media; 2002.
37. Maturana HR, Varela FJ. Autopoiesis and Cognition: The Realization of the Living. vol. 42 of Boston Studies in the Philosophy and History of Science. Dordrecht: Springer Netherlands; 1980.
38. Villalobos M, Dewhurst J. Enactive autonomy in computational systems. Synthese. 2018; 195(5):1891–1908. https://doi.org/10.1007/s11229-017-1386-z
39. Langton CG. Computation at the Edge of Chaos: Phase Transitions and Emergent Computation. Physica D: Nonlinear Phenomena. 1990; 42(1-3):12–37. https://doi.org/10.1016/0167-2789(90)90064-V
40. Zenil H, Kiani N, Tegnér J. A Review of Graph and Network Complexity from an Algorithmic Information Perspective. Entropy. 2018; 20(8):551. https://doi.org/10.3390/e20080551 PMID: 33265640
41. Barabási AL. Network Science. 1st ed. USA: Cambridge University Press; 2016.
42. Abrahão FS, Wehmuth K, Ziviani A. Algorithmic networks: Central time to trigger expected emergent open-endedness. Theoretical Computer Science. 2019; 785:83–116. https://doi.org/10.1016/j.tcs.2019.03.008
43. Abrahão FS, Wehmuth K, Ziviani A. Emergent Open-Endedness from Contagion of the Fittest. Complex Systems. 2018; 27(04).
44. University of Glasgow. Assembly Theory Unifies Physics And Biology To Explain Evolution And Complexity. Press Release. 2023;.

45. Ivanciuc O. Chemical Graphs, Molecular Matrices and Topological Indices in Chemoinformatics and Quantitative Structure-Activity Relationships. Current Computer Aided-Drug Design. 2013; 9(2):153–163. https://doi.org/10.2174/1573409911309020002 PMID: 23701000

46. Mowshowitz A, Dehmer M. Entropy and the complexity of graphs revisited. Entropy. 2012; 14(3):559–570. https://doi.org/10.3390/e14030559

47. 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. https://doi.org/10.1007/s00239-017-9824-6 PMID: 29260254

48. Ziv J, Lempel A. A Universal Algorithm for Sequential Data Compression. IEEE Transactions on Information Theory. 1977; 23(3):337–343. https://doi.org/10.1109/TIT.1977.1055714

49. Ziv J, Lempel A. Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory. 1978; 24:530–536. https://doi.org/10.1109/TIT.1978.1055911

50. Zenil H, Soler Toscano F, Gauvrit N. Methods and Applications of Algorithmic Complexity: Beyond Statistical Lossless Compression: 44 (Emergence, Complexity and Computation, 44). Berlin, Heidelberg: Springer; 2022.

51. Zenil H, Delahaye JP, Gaucherel C. Image Characterization and Classification by Physical Complexity. Complexity. 2012; 17(3):26–42. https://doi.org/10.1002/cplx.20388

52. Zenil H, Adams A, Abrahão FS. Optimal Spatial Deconvolution and Message Reconstruction from a Large Generative Model of Models. arXiv Preprints. 2023;(arXiv:1802.05843).

53. Ozelim L, Uthamacumaran A, Abrahão FS, Hernández-Orozco S, Kiani NA, Tegnér J, et al. Assembly Theory Reduced to Shannon Entropy and Rendered Redundant by Naive Statistical Algorithms. arXiv Preprints. 2024;(arXiv:2408.15108).

54. Marshall SM, Moore DG, Murray ARG, Walker SI, Cronin L. Formalising the Pathways to Life Using Assembly Spaces. Entropy. 2022; 24(7):884. https://doi.org/10.3390/e24070884 PMID: 35885107

55. Zenil H. Lee Cronin's Assembly Theory Disputed & Debunked by Dr. Hector Zenil. Youtube; 2023. Available from: https://www.youtube.com/watch?v=078EXZeS8Y0&.

56. Marshall SM, Moore D, Murray ARG, Walker SI, Cronin L. Quantifying the pathways to life using assembly spaces. arXiv Preprints. 2019;.

57. Ming Li, Xin Chen, Xin Li, Bin Ma, Vitanyi P. Clustering by Compression. IEEE International Symposium on Information Theory, 2003 Proceedings. 2003; p. 261–261.

58. Dauwels J, Srinivasan K, Ramasubba Reddy M, Musha T, Vialatte FB, Latchoumane C, et al. Slowing and Loss of Complexity in Alzheimer's EEG: Two Sides of the Same Coin? International Journal of Alzheimer's Disease. 2011; 2011(1):539621. https://doi.org/10.4061/2011/539621 PMID: 21584257

59. Lempel A, Ziv J. On the complexity of finite sequences. IEEE Transactions on information theory. 1976; 22(1):75–81. https://doi.org/10.1109/TIT.1976.1055501

60. Cover TM, Thomas JA. Elements of Information Theory. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2005.

61. Von Korff M, Sander T. Molecular Complexity Calculated by Fractal Dimension. Scientific Reports. 2019; 9(1):967. https://doi.org/10.1038/s41598-018-37253-8 PMID: 30700728

62. Delahaye JP, Zenil H. Numerical evaluation of algorithmic complexity for short strings: A glance into the innermost structure of randomness. Applied Mathematics and Computation. 2012; 219(1):63–77. https://doi.org/10.1016/j.amc.2011.10.006

63. Krakauer DC, Plotkin JB. Redundancy, antiredundancy, and the robustness of genomes. Proceedings of the National Academy of Sciences. 2002; 99(3):1405–1409. https://doi.org/10.1073/pnas.032668599 PMID: 11818563

64. 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. https://doi.org/10.1098/rsta.2016.0342 PMID: 29133442

65. Bar-Yam Y. A mathematical theory of strong emergence using multiscale variety. Complexity. 2004; 9(6):15–24. https://doi.org/10.1002/cplx.20029

66. Auerbach JE, Bongard JC. Environmental influence on the evolution of morphological complexity in machines. PLoS computational biology. 2014; 10(1):e1003399. https://doi.org/10.1371/journal.pcbi.1003399 PMID: 24391483

67. Abrahão FS, Wehmuth K, D'Ottaviano IML, Carvalho LLd, Zenil H. Expected emergent open-endedness from partial structures extensions under algorithmic perturbations. In: Theoretical and Foundational Problems in Information Studies (TFPIS); 2021. Available from: https://tfpis.com/.

68. Jaeger J. Assembly Theory: What It Does and What It Does Not Do. Journal of Molecular Evolution. 2024. https://doi.org/10.1007/s00239-024-10163-2 PMID: 38453740

69. Hernández-Orozco S, Hernández-Quiroz F, Zenil H. The Limits of Decidable States on Open-Ended Evolution and Emergence. ALIFE 2016, the Fifteenth International Conference on the Synthesis and Simulation of Living Systems. 2016; p. 200–207.

70. Kirchherr W, Li M, Vitányi P. The Miraculous Universal Distribution. The Mathematical Intelligencer. 1997; 19:7–15. https://doi.org/10.1007/BF03024407

71. Festival WS. The Limits of Understanding; 2015. YouTube. Available from: https://www.youtube.com/watch?v=DfY-DRsE86s&t=5392s.