

# The determinism and boundedness of self-assembling structures

S. Tesoro,<sup>1</sup> S. E. Ahnert,<sup>1,2</sup> and A. S. Leonard<sup>1,2,\*</sup>

<sup>1</sup>*Theory of Condensed Matter, Cavendish Laboratory,  
University of Cambridge, CB3 0HE Cambridge, UK*

<sup>2</sup>*Sainsbury Laboratory, University of Cambridge, CB2 1LR Cambridge, UK*

Self-assembly processes are widespread in nature, and lie at the heart of many biological and physical phenomena. The characteristics of self-assembly building blocks determine the structures that they form. Two crucial properties are the determinism and boundedness of the self-assembly. The former tells us whether the same set of building blocks always generates the same structure, and the latter whether it grows indefinitely. These properties are highly relevant in the context of protein structures, as the difference between deterministic protein self-assembly and nondeterministic protein aggregation is central to a number of diseases. Here we introduce a graph theoretical approach that can determine the determinism and boundedness for several geometries and dimensionalities of self-assembly more accurately and quickly than conventional methods. We apply this methodology to a previously studied lattice self-assembly model and discuss generalizations to a wide range of other self-assembling systems.

## I. INTRODUCTION

Self-assembly is a ubiquitous phenomenon in nature, producing complex structures in biology, chemistry and physics. Examples include DNA [1–4], protein quaternary structure [5, 6], protein aggregation [7], viruses [8], micelles [9], and thin films [10].

Two fundamental questions about a self-assembling system are: do the building blocks always form the same structure and does the assembly grow indefinitely. The former property is referred to as the *determinism* and the latter as *boundedness* [11, 12]. Together they form the assembly classification.

Protein complexes are a prominent example of deterministic, bound self-assembly. Misfolding or erroneous binding of mutated versions of such proteins can cause the self-assembly of a protein complex to become nondeterministic, unbound protein aggregation. This in turn is the hallmark of a number of severe diseases, such as sickle-cell anemia and Alzheimer’s [7, 13].

In this paper we introduce a framework for establishing the determinism and boundedness of a given set of self-assembling building blocks. We apply this approach to a previously studied lattice self-assembly model introduced in [14] and studied further in [11, 12, 15–17]. In this model square tiles have attractive interfaces of different types, with interactions governed by a simple set of rules. The final structures assembled are sets of connected lattice sites known as polyominoes.

Such tile self-assembly or polyomino models have been useful for the study of genotype-phenotype maps, where the specification of the building blocks can be viewed as a genotype and the resulting structure as the phenotype [11, 15, 16]. This approach can be combined with genetic algorithms to model evolutionary processes [11].

Despite being an abstract model, this polyomino model can directly and meaningfully map to real biological self-assembly phenomena. For example, the sickle-cell mutation of hemoglobin that leads to unbound protein aggregation can be modelled using polyominoes [15]. Furthermore, experimental implementations of the polyomino model have been realized using DNA tiles [17].

The assembly process previously used is fully stochastic, and may be sketched out as: (a) the structure is seeded with a randomly selected tile, (b) a random face on the structure is chosen, (c) a random tile is drawn with a random orientation, and (d) the drawn tile binds to the structure if the interfaces are interacting. Steps (b-d) are then repeated until no further attachments are possible and assembly terminates.

In this paper, we introduce a graph based approach to replace stochastic assembly as the primary method for identifying the determinism and boundedness of a given set of self-assembly building blocks. The stochastic approach is inelegant and suffers from a compromise between accuracy and speed, whereas the graph theoretical approach offers a robust methodology that improves accuracy and speed.

This paper proceeds by discussing self-assembly basics in Section II. Assembly graphs and their construction from tile sets are introduced in Section III. Classification preserving transformations of assembly graphs which simplify graph features into deterministic motifs are discussed in Section IV. After classifying the assembly graph, the structure is validated under steric constraints in Section V. Extensions to the method and general discussion are in Sections VI and VII respectively.

With generalizations to other geometries and dimensions, there are potential applications in the study of protein complexes and protein aggregation, as well as bioengineering and nanotechnology.

---

\*Electronic address: asl47@cam.ac.uk

## II. LATTICE SELF-ASSEMBLY

Following the model introduced in [14], a lattice self-assembly tile set consists of one or more tile types. Each side of a tile (the geometric face) has an interface type, with every tile type in the set exhibiting unique configurations of these interfaces. Conventionally, interactions are defined with  $\mathbf{0}$  as non-interacting and  $\mathbf{1} \leftrightarrow \mathbf{2}$ ,  $\mathbf{3} \leftrightarrow \mathbf{4}$ , etc. as interacting pairs. However, these interaction rules can take any arbitrarily complex form provided interactions are bidirectional. Interactions are infinite in strength, meaning two tiles bind irreversibly if the adjoined interfaces are interacting. There is an infinite population of each tile type, precluding any stoichiometric limitations.

### A. Definitions of structure and determinism

A structure is defined as a set of connected tiles, each with an associated tile type, orientation, and unique lattice coordinates. Structure determinism can be defined in three distinct categories, listed in increasing order of strictness: shape, tile, and orientation determinism. Structures are defined independently of absolute position and rotation, and so lattice coordinate translations or rotations of entire structures are considered indistinguishable (structures are *one-sided* polyominoes). An example of a tile set and its assembled polyomino is shown in Figure 1.

**Shape determinism** requires all produced structures to be the same one-sided polyomino.

**Tile determinism** additionally requires corresponding coordinates between two structures to have matching tile types.

**Orientation determinism** further requires matching relative inter-tile orientations.

Orientation determinism is standard choice due to the impact that tile type and relative orientation could have on the function of the assembled structure.

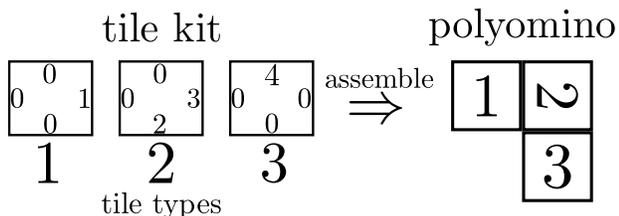


FIG. 1: A tile set (left) and its assembled polyomino (right). Each tile type in the set can be assigned a number, which is used to label the tile type in the polyomino. Rotating the label indicates the corresponding rotation of the tile during assembly.

## III. ASSEMBLY GRAPHS

Assembly tile sets are represented using the notation {tile type 1, tile type 2, ...}, where tile type ordering is irrelevant. Tile types have cyclic symmetry and have a length fixed by the geometry, so square tiles have four faces and can be represented as  $(F_1, F_2, F_3, F_4)$ , where  $F_i$  indicates the interface type of face  $i$ . We use the convention that faces are encoded clockwise starting from the top. An example of this notation can be found in the caption of Figure 2.

The **assembly graph** of a tile set is constructed with nodes for each face on every tile type. The faces within each tile type form cliques with edges labeled as internal, while interactions between interfaces (both inter- and intra-tile) are encoded with edges labeled as external. Hence an assembly graph can be represented using an edge-labeled pseudograph (multigraph with loops). A detailed assembly graph example is shown in Figure 2. Internal edges only depend on geometry, and are not displayed in future examples for clarity.

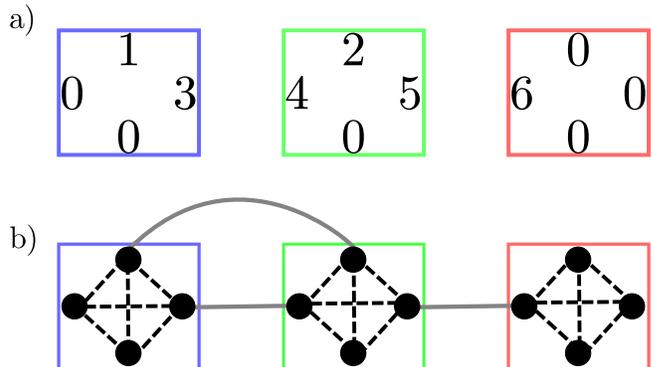


FIG. 2: (a) The building blocks for the tile set  $\{(1, 3, 0, 0), (2, 5, 0, 4), (0, 0, 0, 6)\}$  explicitly labeled with interfaces and the assembled structure. (b) The corresponding complete assembly graph with black (dashed) internal edges, and gray (solid) external edges.

### A. Assembly graph terminology

Assembly graphs contain three features of interest: single interacting faces, branching points, and cycles. Figure 3 shows examples of partial assembly graphs with such features.

**Single interacting face (SIF)** tiles are tile type which have only a single face with one or more external edges.

**Branching points** occur when a given face has multiple external edges. Non-SIF branching points occur if a tile type has a branching point and at least one other face with external edges. Non-SIF branching points frequently cause nondeterminism due to diverging assembly pathways.

**Cycles** have two varieties with identical impact on assembly classification: inter- and intra-tile. Inter-tile cycles are walks on the assembly graph which alternate stepping on external and internal edges. Intra-tile cycles are walks of only a single step on an external edge connecting to the same tile. Cycles are the primary source of unboundedness, due to the potential to endlessly traverse (and thus assemble) around the cycle.

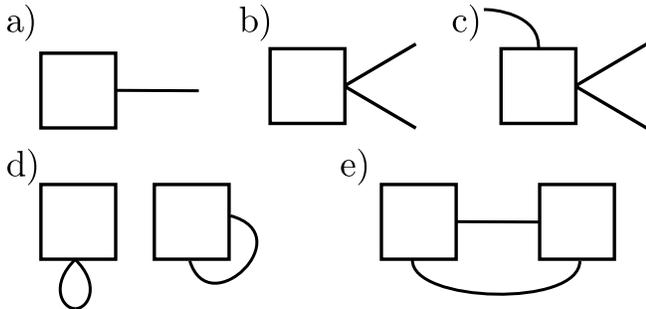


FIG. 3: Five partial assembly graphs demonstrating various features. (a) SIF tile (b) SIF branching point (c) non-SIF branching point (d) two intra-tile cycles (e) inter-tile cycle.

## B. Graph connectivity

If an assembly graph contains disconnected components, it is trivially nondeterministic due to seed dependence. This is due to the inability to assemble between disconnected components (by definition), and hence the structures formed by seeds in different components are nondeterministic.

However, each connected component can be assessed independently for determinism. Biologically, this can be related to genotypes encoding multiple independent phenotypes.

## C. Fundamental Deterministic Graph

A **treelike** assembly graph is one without any cycles or non-SIF branching points. The simplest treelike graph is a single tile with no interactions, and is evidently bound and deterministic. Adding a SIF tile (potentially SIF branching point) with a new interaction pair to this assembly graph cannot introduce a non-SIF branching point or cycle by definition. Since these are the only sources of nondeterminism and unboundedness, the assembly classification consequently cannot be altered. By induction, any treelike graph is therefore bound and deterministic.

As such, any assembly graph that can be transformed into a treelike graph will also be bound and deterministic. We now focus on the procedures of reducing an arbitrarily complex assembly graph to this deterministic

state, or if this is not possible, identifying the source of nondeterminism or unboundedness.

Sequentially, we prune SIF tiles and check if the assembly graph is treelike. If we cannot classify the assembly graph at this stage, we analyze the nature of any cycles present and remove trivial cycles, and examine the remaining cycles for unbound behavior. Although the graph approach identifies sources of nondeterminism and unboundedness, it is unable to predict spatial conflicts that lead to steric nondeterminism, a limitation discussed later.

## IV. ALTERNATIVE TREELIKE ASSEMBLY PROCEDURES

Explicitly transforming assembly graphs to be treelike is unnecessary; the steps of SIF tile pruning to remove trivial branching points and checking for infinite cycles is sufficient to determine the assembly classification.

Assembly graphs which cannot undergo or fail the above procedures are necessarily nondeterministic or unbound. For instance, any assembly graph with multiple of each complementary interface, i.e. branching points connected to branching points, cannot be made treelike and thus is nondeterministic.

### A. SIF tiles elimination procedure

SIF tiles, by converse reasoning to that of the fundamental deterministic graph in Section III C, can be ‘pruned’ from an assembly graph without altering its assembly classification. This is done by removing a SIF tile from the assembly graph, and neutralizing its complementary interface (or interfaces if it’s a SIF branching point). SIF tiles whose complementary interface is a branching point are nondeterministic and may be classified without pruning.

This procedure is applied iteratively, and if a treelike graph is obtained after a removal, the assembly graph is known to be bound and deterministic. Examples of the procedure is shown in Figure 4.

### B. Cycle rank classification

Cycles can be categorized by the number of times the cycle pattern is repeated during the assembly, a quantity known as the rank. The rank of the cycle can be determined from a single traversal of the cycle and noting the net rotation accumulated at the end of the walk. Figure 5 shows several examples of cycles and their rank classification.

For square geometry, there are 4 periodic net rotations to consider,  $\theta = n\pi/2$  with  $n \in \{-1, 0, 1, 2\}$ . Rank 4 cycles occur from  $|n| = 1$ , while  $n = 2$  produces rank 2

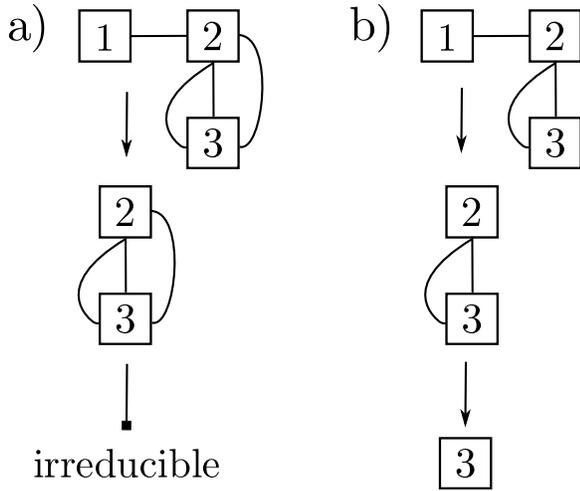


FIG. 4: The SIF elimination procedure applied iteratively on two assembly graphs. The ‘1’ SIF tile may be pruned in both (a) and (b), but only in (b) does the branching point become a SIF and allow further pruning. Note that the second stage of (b) is already treelike (no cycles or non-SIF branching points) and thus classified. Since (a) cannot be reduced to treelike, the assembly graph cannot be classified yet.

cycles. There are two subcategories for  $n = 0$ , rank  $\infty$  and rank 1, based on spatial considerations.

If after traversing the cycle once, there is a zero net spatial translation, i.e. the final tile placed reattaches in real-space to the first tile, then the cycle is rank 1. By contrast, if there is a nonzero net spatial translation and the final tile placed leaves the final interface exposed, then the cycle will grow ad infinitum and is rank  $\infty$ .

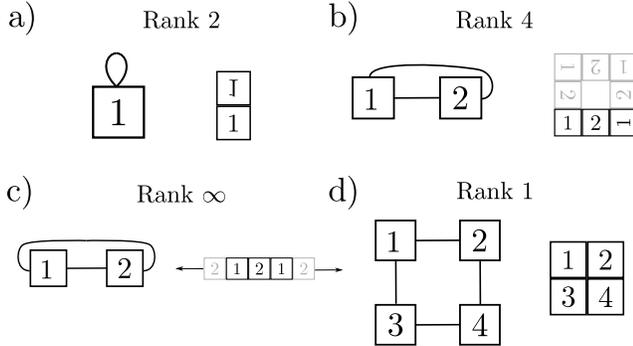


FIG. 5: The four cycle ranks possible with square geometry with assembly graphs on the left and assembled structures on the right in each example. (a) rank 2 intra-tile cycle ( $\Delta\theta = \pi$ ) (b) rank 4 inter-tile cycle ( $\Delta\theta = -\pi/2$ ) (c) rank  $\infty$  inter-tile cycle ( $\Delta\theta = \infty, \Delta_{XY} \neq 0$ ) (d) rank 1 inter-tile cycle ( $\Delta\theta = \infty, \Delta_{XY} = 0$ ). As soon as the initially placed tile is reused, the cycle rank may be classified, but for completeness the remaining assembly is shown in gray.

### C. Establishing boundedness

Cycles of rank 1, like SIFs, contribute trivial assembly behavior, and can be simplified without interfering with assembly classification. Any edge in the cycle may be removed, even if an edge is shared with another cycle. Regardless of the choice of cut, the resulting assemblies graphs will have the same quantity and rank of surviving cycles.

This procedure is likewise applied iteratively along with SIF elimination until the assembly graph is maximally simplified. More details are given in Appendix A. If at any stage a rank  $\infty$  cycle is discovered, growth is immediately known to be unbound. Moreover, multiple surviving cycles ordinarily exhibit unboundness due to their amalgamated assembly pattern.

Simplistically, this can be understood as the regeneration of cycle interfaces. When one cycle completes assembling, interfaces that assemble the other cycles are exposed and available for further growth. This is demonstrated in Figure 6.

Multiple surviving cycles can have rank 1 behavior rather than rank  $\infty$ , although it’s rare due to the specific spatial constraints required. This is detailed in Appendix A for the specific case of Figure 6.

Establishing the (in)finite behavior of the surviving cycles is hence nuanced, and proper care must be taken to identify infinite cycle behavior. Fundamentally, infinite behavior can be identified if a tile in a cycle of finite rank  $R$  is ever used in assembly more than  $R$  times.

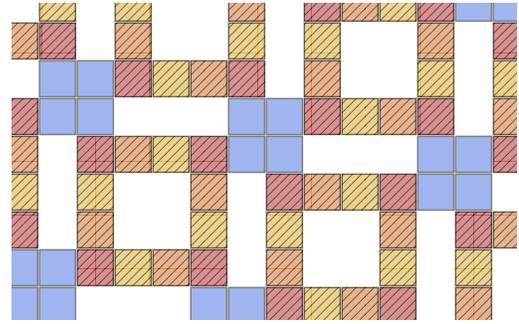


FIG. 6: A section of an unbound structure resulting from an assembly graph with two rank 4 cycles. Every time the blue (plain) cycle is completed, faces which interact with the red-hued (hatched) cycle are left exposed. A completion of the red-hued cycle likewise allows new copies of the blue cycle to start assembling, leading to unbound growth.

## V. COMPLETE PROCEDURE

### A. Steric effects

In the preceding sections we have addressed whether the assembly interactions alone make a tile set (non)deterministic or (un)bound. However, steric effects

can also impact the growth of structures and cause nondeterminism and boundedness in assembly graphs that are otherwise deterministic or unbound.

For this reason a steric validation must be performed by growing the structure on the real-space lattice once. If any lattice coordinate is used in assembly multiple times, then the assembly is sterically nondeterministic. Such steric nondeterminism is clearly illustrated in Figure 7.

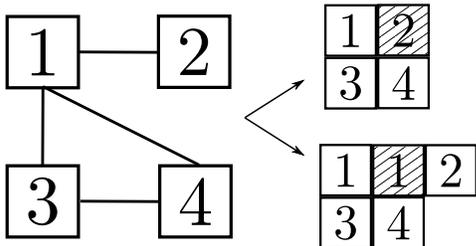


FIG. 7: This assembly graph is rule-deterministic, but fails steric validation. Growing the structure on the real-space lattice reveals steric nondeterminism as the hatched lattice site does not have a unique occupant. This leads to nondeterminism as the final structure would depend on which occupant was built first.

Steric nondeterminism can prompt novel behavior in self-limiting cluster growth. In previous experimental work [17], single-seed and multi-seed self-assembly were compared, observing that complementary pair interactions can limit growth in multi-seed assembly through local steric effects. Such phenomenon is beyond the scope of this framework currently, but remains an topic for further analysis.

### B. Geometric symmetry

The presence of symmetric tiles will never impact classifying a bound and deterministic assembly, but may mistake unbound growth for nondeterminism due to the symmetry-induced branching points. These spurious nondeterministic branching points can be replaced with alternative interfaces that preserve assembly classification through a desymmetrization procedure (see Appendix C for details).

### C. Analysis Sequence

The flowchart in Figure 8 illustrates the steps in determining the (un)bound and (non)deterministic nature of a self-assembling tile set. An open-source implementation of the assembly graph and steric algorithms is available online[18].

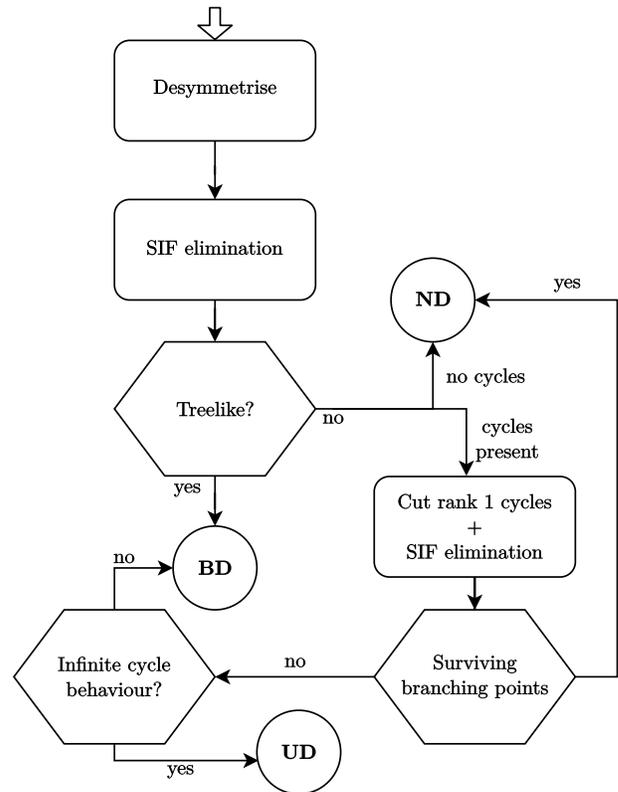


FIG. 8: The sequential analysis of an assembly graph. Classifications can be bound and rule-deterministic (BD), unbound rule-deterministic (UD), or nondeterministic (ND). Steric validation must still be performed to ensure determinism. Note disconnected graphs are trivially seed-dependent, but each connected component may undergo this analysis.

## VI. FRAMEWORK EXTENSIONS

The modular nature of the assembly graph algorithms allows the pinpointing of unbounded and nondeterministic behavior. Relaxing the conditions which detect the above classifications allows examining a wider range of assembly conditions and constraints.

### A. Other dimensions and geometries

The procedures introduced here directly extend to regular triangular and hexagonal geometries, which allow regular tilings of the plane[19]. The only conceptual modification is the  $\theta$ -shifts in cycle rank classification. Triangular geometry allows finite cycles of rank 1, 2, and 6, while hexagonal geometry allow ranks of 1, 2, 3, 4, and 6. Some cycles examples for regular hexagons are presented in Figure 9.

The procedures can also be extended to other dimensions, again only modifying cycle rank classification. For instance, cubic geometry still only supports finite cycles ranks of 1, 2, and 4, while in one dimension only rank 2

finite cycles are possible. Although identifying infinite cycle behavior, steric validation, and other mentioned procedures are more complicated to implement, the concepts are unaltered.

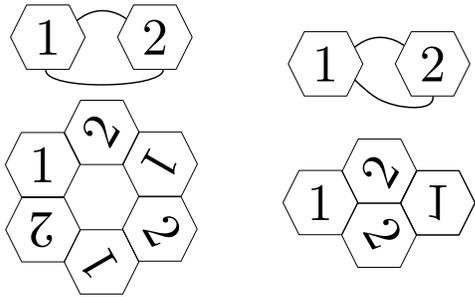


FIG. 9: Hexagonal geometry allows for additional cycle ranks. A rank 3 and rank 2 cycle and the resulting structures are shown on the left and right respectively.

### B. Nonstatic interaction rules

Various biological assemblies do not follow the earlier definition of determinism, due to the ability to form different final structures depending on external conditions, known as *phenotype plasticity*. Such external conditions can considerably impact self-assembly of biomolecules [8, 20]. This framework can encapsulate plasticity by considering e.g. high pH interaction rules and low pH interaction rules, where the choice of rules determines which interactions are active. This in turn can yield different deterministic structures.

Likewise, it is possible to consider sequentially active interaction rules in a single assembly. An assembly graph constructed under each set of interaction rules can be analyzed for (un)boundness and (non)determinism. The assembly is then validated under steric constraints by assembling under the first set of interaction rules, followed by using the next set of rules on the existing structure etc. As such, this approach can generalize to more complex assembly environments.

### C. Seed dependence

Although seed independence is ingrained in the method as outlined, incorporating fixed seed assembly allows for greater diversity in assembled structures. Seed dependence primarily results from branching points, as the behavior of SIF tiles and cycles are independent of the order of assembly.

One possible realization for modifying seed dependence is to walk on the assembly graph depth-first away from the seed. Branching points where a “branched to” face (i.e. multiple faces with this interface type exist in the set) is discovered before the complementary “branching”

face (i.e. only face with this interface type in the set) can be ignored. Intuitively this removes the nondeterminism, as the assembly never encounters the multiple pathways when growing from those seeds. An illustration of this extension is shown in Figure 10.

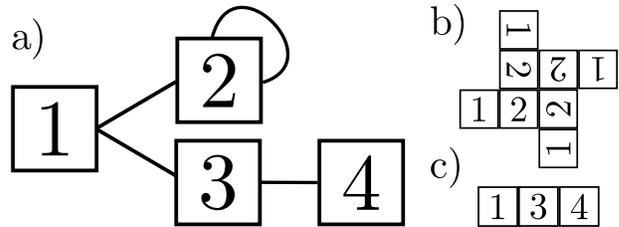


FIG. 10: An assembly graph that is classified as nondeterministic. However, if the assembly is seeded with tile type ‘2’ or ‘3’/‘4’, the structures are deterministic. The nondeterministic behavior of the branching point is only encountered walking from the ‘1’ seed.

## VII. DISCUSSION

### A. Advantages of the assembly graph framework

The stochastic assembly method suffers from the probabilistic detection of unboundness and nondeterminism. Classification accuracy is directly related to the number of repeated assemblies,  $K$ , which compromises speed. For example, stochastic assembly can miss the unbounded and nondeterministic nature of the tile set  $\{(1, 0, 2, 0), (1, 2, 0, 0)\}$  even with seed independence. The misclassification probability for this tile set has an analytic form for  $K$  repeated assemblies given by

$$\text{Pr} = \sum_{N=1}^{\infty} [N(1/2)^{3+2N}]^K$$

where  $N$  is the number of pairs of the infinite cycle  $(1, 0, 2, 0)$  tiles in the assembly. For  $K = 10$ , the probability of misclassification is inconsequential ( $10^{-15}$ ). However, more complex tile sets or weaker determinism definitions can lead to non-negligible loss of accuracy. Assembling  $\{(1, 2, 0, 0), (1, 0, 0, 0)\}$  with the less restrictive shape determinism has a misclassification rate of  $(7/8)^{K/2}$ , over 50% for  $K = 10$ .

The form of stochastic assembly algorithm (an example of which is sketched out in Section I) also impacts probabilistic detection. In the immediately preceding tile set example, randomly drawing a new tile gives equal probability to the two tile types. However, the  $(1, 2, 0, 0)$  tile type has two potential bindings, and so could also be considered to have twice the probability to be selected over the other tile type. Depending on the implementation choice, the misclassification rate varies between  $(19/27)^{K/2}$  and  $(7/8)^{K/2}$ . The graph method removes

the need for external parameters and model specifications inherent in stochastic assembly.

Due to the exponential growth of configuration space, only one, two, and three tile systems have been exhaustively searched. However, assembly graphs up to 20 tiles were tested with  $10^{11}$  samples per graph size ( $N_T$ ). All sampled tile sets exhibited no misclassifications when compared with the “truth” outcomes of sufficiently reliable ( $K = 50N_T$ ) stochastic assemblies. A direct comparison of the two methods’ speeds is shown in Figure 11 for connected assembly graphs. Details on the simulations used can be found in Appendix B.

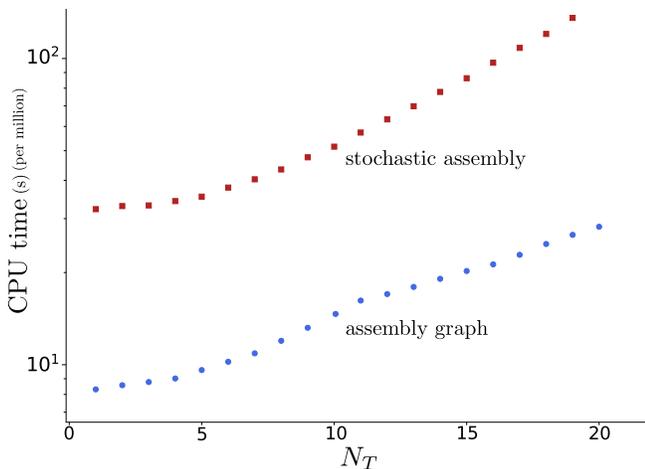


FIG. 11: Elapsed CPU time (s) per million connected assembly graphs classified for the graph approach and stochastic assembly with  $K = 10$ . Both methods initially scale with  $N_T$  similarly, although with the graph approach approximately 3 times faster. However, for larger  $N_T$  the stochastic method scales superlinearly while the graph framework scaling remains approximately linear.

### B. Nondeterministic assembly dynamics

In [12], nondeterministic self-assembly of single lattice tiles and mixtures of two tiles at varying relative concentrations were studied in detail. In addition to using the conventional interaction rules introduced initially, assembly was also examined using **1s** and **2s** as self-interacting. Tile sets were classified in terms of their behavior upon variation of tile density, recognizing critical transitions from bound to unbound growth and noncritical density transitions.

These dynamics can be analyzed from a different perspective, examining the 106 topologically distinct tile sets with the assembly graph framework. While identifying deterministic assemblies is relatively unchanged, the ability to robustly distinguish unbound from nondeterministic behavior is greatly improved using the assembly graph framework.

### C. Genotype-phenotype maps

The lattice self-assembly model described here has been used as an abstract model of the genotype-phenotype (GP) map of protein quaternary structure [15]. In these models, tile set is the genotype and the assembled structure represents the phenotype. This work focused on bound deterministic assembly, while unbound or nondeterministic building block sets were regarded as a single unfavorable phenotype, and largely ignored.

The assembly graph framework allows assembly graphs to be used as an intermediate link between genotypes and phenotypes, and thus extend such work to examine the GP map of nondeterministic tile sets.

### D. Application to real proteins

The study of such nondeterministic GP maps is essential, as nondeterministic self-assembly in the form of protein aggregation is a hallmark of numerous diseases. A classic example is that of hemoglobin and the sickle cell mutant.

In this representation, the  $\alpha$  and  $\beta$  chains of wild-type hemoglobin Hb A can be mapped to the  $\{(1, 3, 0, 0), (2, 0, 0, 4)\}$  tile set, and the sickle-cell mutant Hb S can be mapped to  $\{(1, 3, 0, 0), (2, 5, 6, 4)\}$ , displayed in Figure 12. The sickle-cell point mutation is sufficient to introduce an interaction labeled here with 5 and 6. The wild-type assembly graph possesses a cycle of rank 2, and the mutation in the second tile introduces a new cycle of rank 4. Following the methodology we have outlined, this is sufficient to identify the unbound deterministic growth giving rise to the sickle cell anemia disease.

Biological structures that are assembled through cycles with rank greater than 1 are prone to unbound growth via mutations which introduce additional cycles. Protein misfolding and resulting changes of interface angles in quaternary structures can be recast as introductions of cycles or branching points.

While many proteins obviously exhibit more complex geometries than this framework is suited to, examining coarse-grained quaternary structure has already revealed insights into assembly properties and evolution history of proteins [5, 21]. Combining these protein “motifs” with the introduced assembly graph formalisms is a rich vein for future exploration.

### E. Universal computing

The motivations behind this method are unassociated with those of universal computing and Turing machines, but there is conceptual overlap with other tile assembly models (TAMs). The interaction rules used here define *noncooperative* bindings, meaning the interaction is independent of factors beyond the two adjoined faces. Such noncooperative models have been demonstrated to

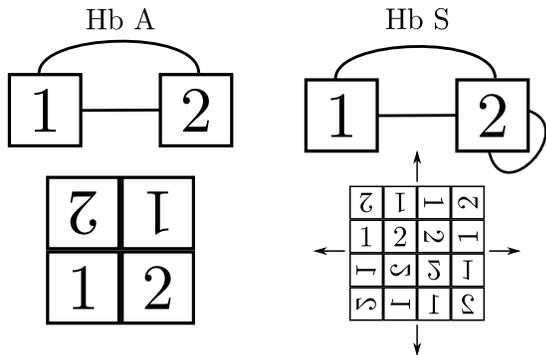


FIG. 12: The wild-type (left) and sickle-cell mutant (right) of the hemoglobin complex has been modelled previously using polyominoes[15]. The unbound growth denoted by arrows resulting from the mutation can now succinctly be pinpointed to the addition of a higher order cycle to the assembly graph.

not intrinsically allow universal computing[22], requiring significant generalizations to the TAMs to increase their computational ability[23].

As such, the framework outlined here for determining boundedness and determinism of self-assembling tile sets has little bearing on universal computation and the infamous halting problem.

## VIII. CONCLUSION

With the assembly graph framework we have introduced a new approach that can determine the (un)boundedness and (non)determinism of self-assembling tile sets. While the results presented have focused on 2D square geometry, the assembly framework readily extends to other dimensionalities (1D or 3D) and other geometries which regularly tile the plane.

The graph based approach outperforms the existing stochastic assembly approach in both speed and accuracy, facilitating the study of significantly larger GP maps and evolutionary dynamics. External parameter dependencies, like the  $K$  repeated assemblies, are also eliminated.

The topological nature of the assembly graphs also opens new options for describing complexity and other properties relevant to evolutionary dynamics. In addition, this methodology is a powerful tool for generalizing the study of GP maps by extending to the nondeterministic realm, which yields a coarse-grained model relevant to dysfunction and disease in the context of biological self-assembly.

## Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (ST and ASL), the

Gatsby Foundation (ASL and SEA), and the Royal Society (SEA). The authors would like to thank the referees for extensive feedback and suggestions on this work.

## Appendix A: Cycles

Determining the rank of a cycle can be done with a single walk around the cycle and tracking the net rotation and translation. Since each tile has geometry specific cyclic symmetry, the individual steps on the walk contain no information on cycle rank; only the whole walk is useful. Cycle walks that encounter branching points are almost strictly nondeterministic. For completeness, these walks can proceed normally, continuing an independent “cloned” walk down each branched edge.

There are combinatorial arguments that cycles must have necessarily formed once there are more edges than tiles (accounting for branching points), and the cycles can be detected in many ways. Exhaustive walks, effectively a depth-first search, on the graph will detect all cycles, or more advanced algorithms can be adapted like Prim’s or Kruskal’s.

### 1. Cycle cutting

As soon as a rank 1 cycle is discovered, any single edge in the cycle may be cut without impacting assembly classification. This is due to the fact that final step in the cycle is unnecessary: the walk returns to the starting point with no net rotation. An example is shown in Figure 13. While the choice of edge doesn’t change the assembly classification, there often is an “optimal” choice of cut. In Figure 13, cutting ‘A’ leaves a larger assembly graph, which slows analysis. However, determining the optimal cut typically involves a similar number of operations as handling the less optimal cut, and so offers little performance gain.

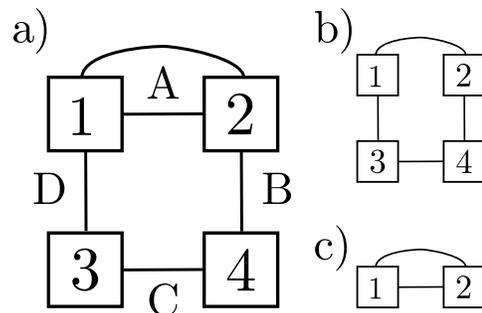


FIG. 13: The assembly graph in (a) contains a cycle of rank 1 and one of rank 2. Cutting edge A results in the (b) assembly graph, while cutting edge B, C, or D results in the (c) assembly graph. Any choice of cut results in an assembly graph with a single rank 2 cycle.

## 2. Nested cycles

In general, multiple surviving cycles will grow infinitely. However, if the multiple cycles are “nested”, and thus do not regenerate exposed interfaces of the other cycles, then the assembly is bounded.

In the case of Figure 6, this nested behavior can be easily seen by two changes to the tile set. Relocating where the two cycles connect in the assembly graph, changing  $\{(1, 0, 3, 2), (4, 5, 7, 0), (0, 9, 0, 6), (0, 8, 0, 10)\}$  to  $\{(1, 0, 3, 2), (0, 5, 7, 0), (0, 9, 4, 6), (0, 8, 0, 10)\}$ , demonstrates multiple cycles producing rank 1 behavior. This is shown in Figure 14. In this case, the completion of one cycle does not generate new exposed faces for the other cycle, and hence precludes infinite growth.

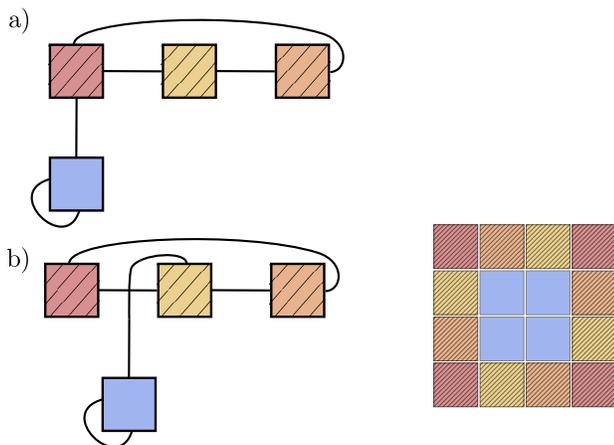


FIG. 14: The assembly graph for Figure 6 is shown in (a), while (b) is a similar assembly graph, but one that is bounded with nested cycle behavior. The only difference in assembly graphs is where the tile with the intra-tile cycle joins the other cycle. The bounded structure for (b) is shown on the right.

## Appendix B: Method comparisons

The connected assembly graph simulations were comprised of  $10^{11}$  samples per graph size ( $N_T$ ), where all sampled assembly graphs had only a single connected component. This constraint was chosen to prevent e.g. a 7 tile graph actually being comprised of 4 and 3 tile disconnected subgraphs. The number of possible interfaces scaled with tiles as  $4N_T + 2$  to ensure every possible assembly graph could be sampled, without inflating the neutral space unnecessarily.

The fraction of configuration space which produces unbound or nondeterministic structures increases with the number of tiles. As such, the accuracy gain drops as the graph size grows since stochastic assembly is able to correctly reject these strongly nondeterministic assembly graphs. However, the assembly graph approach is still faster and never misclassifies at any  $N_T$ .

Evolutionary dynamics from previous work with this model[11] were also tested using both methods. A population of 500 genotypes was evolved under mutation and fitness proportional selection, where the fitness is the size of the assembled structure. Genotypes were allowed to be disconnected, as each component can be analyzed independently ( $K$  assemblies per component), with the maximum fitness across components taken if all were bound and deterministic.

These evolutionary dynamics are most interesting for large configuration spaces (predominantly from large  $N_T$ ). Since stochastic assembly depends on  $K/N_T \gg 1$  for acceptable accuracy, these simulations were either unnecessarily slow for small connected components or inaccurate for large ones. The approach introduced here has no fixed external parameters, and so maintains a fast and accurate analysis that can scale with assembly graph size.

Other simulations were tested, like the nondeterministic assembly dynamics[12], and had similar behavior to the displayed results with a factor of 5 speedup and  $10^{-6}$  accuracy improvement.

## Appendix C: Symmetry removal

Tiles with geometric symmetry can be substituted with alternative tiles which eliminate symmetry induced branching points but preserve the assembly classification. For square geometry, symmetric tiles can take one of four forms:  $(A, A, A, A)$ ,  $(A, A', A, A')$ ,  $(A, B, A, B)$ , and  $(A, 0, A, 0)$ . Here  $A$  interacts with  $A'$  and  $B$  has its complementary interface elsewhere in the tile set (not shown).

Many combinations of symmetric tiles are intrinsically nondeterministic, even without considering symmetry induced branching points, and so are not desymmetrized as the analysis will correctly classify these tile sets.

The type  $(A, A', A, A')$  can be replaced with  $(A, A', B, B')$ , although this is only useful for a single tile set. Otherwise, it is nondeterministic if  $A$  or  $A'$  appear elsewhere in the assembly, and seed-dependent if they don't.

The combination  $\{(A, B, A, B), (A', B', A', B')\}$  assembles identically to  $\{(A, A, A, A), (A', A', A', A')\}$  which was already treated in the main text as  $\{(A, B, C, D), (A', B', C', D')\}$ . This is clearly seen in the tile set e.g.  $\{(1, 1, 1, 1), (2, 2, 2, 2)\}$ , which is transformed to tile set  $\{(1, 3, 5, 7), (2, 4, 6, 8)\}$  in Figure 15. This case can be ignored if there are more than two tiles in the set as the classification is nondeterministic or seed-dependent if any of the interfaces do or do not appear elsewhere in the tile set.

The remaining cases we must consider are the  $(A, A, A, A)$  and  $(A, B, A, B)$  forms combined with  $(A', 0, 0, 0)$ ,  $(A', 0, A', 0)$ , and  $(A', A', A', A')$ . Combinations containing  $(A', A', 0, 0)$  and  $(A', A', A', 0)$  are always nondeterministic and are ignored.

The cases  $\{(A, A, A, A), (A', 0, 0, 0)\}$  and

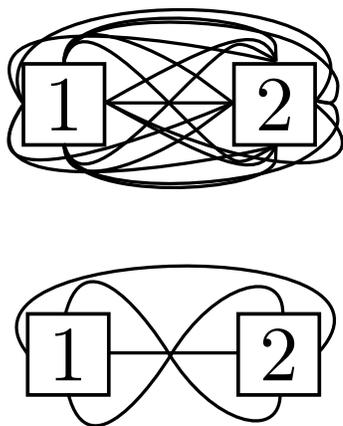


FIG. 15: The desymmetrization procedure removes superfluous branching points arising from symmetry that don't contribute unique assembly behaviour. The top assembly graph is desymmetrized to yield the bottom one, which preserves the unbound deterministic assembly.

$\{(A, B, A, B), (A', 0, 0, 0)\}$  can be ignored as well,

because they are either able to be reduced via the SIF elimination procedure or nondeterministic otherwise. The case  $\{(A, A, A, A), (A', 0, A', 0)\}$  requires the addition of a third tile to preserve classification, and is treated as  $\{(A, B, C, D), (0, D', 0, B'), (0, C', 0, A')\}$ . Finally the case  $\{(A, 0, A, 0), (A', 0, A', 0)\}$  can be replaced with  $\{(A, 0, B, 0), (A', 0, B', 0)\}$ .

In the event of multiple symmetric tiles, the above rules may be combined sequentially to eliminate all symmetry. If an existing symmetry is unable to be eliminated, then the assembly is nondeterministic already as it will strictly not have a treelike expansion. Some of these rules and an example of combined desymmetrization are shown in Figure 16.

In all cases, introduced tiles should be considered “indistinguishable” from the tile it desymmetrizes. This is to prevent any tile or orientation determinism issues during the steric assembly, as they are not actually different tiles but rather a bookkeeping tool. This is shown in Figure 16 by labeling introduced tiles with the original tile label.

- 
- [1] Erik Winfree, Furong Liu, Lisa A Wenzler, and Nadrian C Seeman. Design and self-assembly of two-dimensional DNA crystals. *Nature*, 394(6693):539–544, January 1998.
  - [2] Russell P Goodman, Iwan AT Schaap, Catherine F Tardin, Christoph M Erben, Richard M Berry, Christoph F Schmidt, and Andrew J Turberfield. Rapid chiral assembly of rigid dna building blocks for molecular nanofabrication. *Science*, 310(5754):1661–1665, 2005.
  - [3] Paul WK Rothmund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440(7082):297, 2006.
  - [4] Yonggang Ke, Luvena L Ong, William M Shih, and Peng Yin. Three-dimensional structures self-assembled from dna bricks. *science*, 338(6111):1177–1183, 2012.
  - [5] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann. 3d complex: a structural classification of protein complexes. *PLoS Computational Biology*, 2(11):e155, nov 2006.
  - [6] Gabriel Villar, Alex W Wilber, Alex J Williamson, Parvinder Thiara, Jonathan PK Doye, Ard A Louis, Mara N Jochum, Anna CF Lewis, and Emmanuel D Levy. Self-assembly and evolution of homomeric protein complexes. *Physical review letters*, 102(11):118106, 2009.
  - [7] Samuel IA Cohen, Sara Linse, Leila M Luheshi, Erik Hellstrand, Duncan A White, Luke Rajah, Daniel E Otzen, Michele Vendruscolo, Christopher M Dobson, and Tuomas PJ Knowles. Proliferation of amyloid- $\beta$ 42 aggregates occurs through a secondary nucleation mechanism. *Proceedings of the National Academy of Sciences*, 110(24):9758–9763, 2013.
  - [8] L Lavelle, M Gingery, M Phillips, WM Gelbart, CM Knobler, RD Cadena-Nava, JR Vega-Acosta, LA Pinedo-Torres, and J Ruiz-Garcia. Phase diagram of self-assembled viral capsid protein polymorphs. *The Journal of Physical Chemistry B*, 113(12):3813–3819, 2009.
  - [9] Jacob Israelachvili. Self-assembly in two dimensions: surface micelles and domain formation in monolayers. *Langmuir*, 10(10):3774–3781, 1994.
  - [10] Georg Krausch, Robert Magerle, et al. Nanostructured thin films via self-assembly of block copolymers. *Advanced Materials*, 14(21):1579–1583, 2002.
  - [11] I. G. Johnston, S. E. Ahnert, J. P. K. Doye, and A. A. Louis. Evolutionary dynamics in a simple model of self-assembly. *Phys. Rev. E*, 83:066105, Jun 2011.
  - [12] S. Tesoro and S. E. Ahnert. Nondeterministic self-assembly of two tile types on a lattice. *Phys. Rev. E*, 93:042412, Apr 2016.
  - [13] H F Bunn. Pathogenesis and treatment of sickle cell disease. *The New England Journal of Medicine*, 337(11):762–769, September 1997.
  - [14] S. E. Ahnert, I. G. Johnston, T. M. A. Fink, J. P. K. Doye, and A. A. Louis. Self-assembly, modularity, and physical complexity. *Phys. Rev. E*, 82:026117, Aug 2010.
  - [15] Sam F Greenbury, Iain G Johnston, Ard A Louis, and Sebastian E Ahnert. A tractable genotype–phenotype map modelling the self-assembly of protein quaternary structure. *Journal of The Royal Society Interface*, 11(95):20140249, 2014.
  - [16] Sam F Greenbury, Steffen Schaper, Sebastian E Ahnert, and Ard A Louis. Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLoS Comput Biol*, 12(3):e1004773, 2016.
  - [17] Salvatore Tesoro, K Göpfrich, T Kartanas, Ulrich Felix Keyser, and Sebastian Edmund Ahnert. Nondeterministic self-assembly with asymmetric interactions. *Physical Review E*, 94(2):022404, 2016.
  - [18] A.S. Leonard. Square lattice tile assembly model. <https://github.com/ASLeonard/SLAM>, 2017.
  - [19] Branko Grünbaum and Geoffrey C Shephard. Tilings by regular polygons. *Mathematics Magazine*, 50(5):227–247,

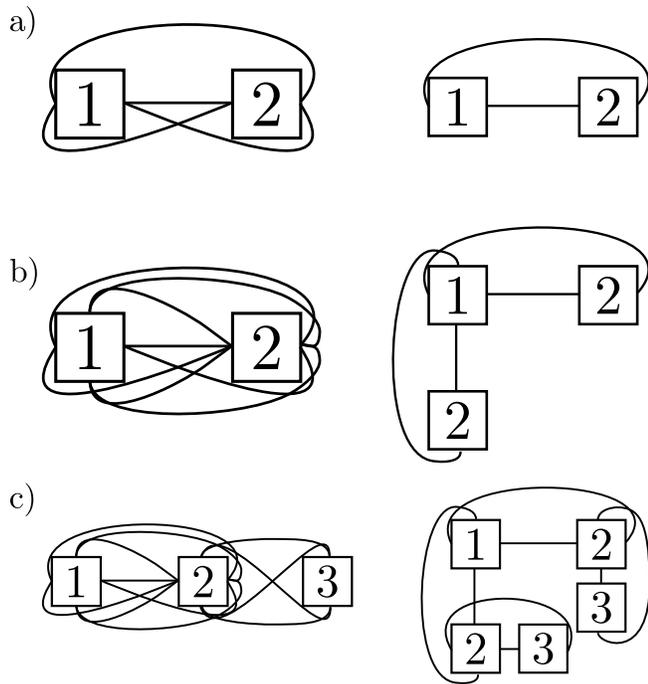


FIG. 16: Three example assembly graphs (left) which can be desymmetrized (right) as described above. (a) and (b) are single desymmetrizations described in the text, while (c) combines both desymmetrization rules used in (a) and (b) into a single assembly graph. Introduced tiles are labeled with the original tile label to ensure the desymmetrized assembly is indiscernible from the original assembly.

- 1977.
- [20] David Deamer, Sara Singaram, Sudha Rajamani, Vladimir Kompanichenko, and Stephen Guggenheim. Self-assembly processes in the prebiotic environment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1474):1809–1818, 2006.
- [21] Sebastian E Ahnert, Joseph A Marsh, Helena Hernández, Carol V Robinson, and Sarah A Teichmann. Principles of assembly reveal a periodic table of protein complexes. *Science*, 350(6266):aaa2245, 2015.
- [22] Pierre-Étienne Meunier and Damien Woods. The non-cooperative tile assembly model is not intrinsically universal or capable of bounded turing machine simulation. *arXiv preprint arXiv:1702.00353*, 2017.
- [23] Jacob Hendricks, Matthew J Patitz, and Trent A Rogers. Doubles and negatives are positive (in self-assembly). *Natural Computing*, 15(1):69–85, 2016.