Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution

Iain G Johnston [1,2], Kamaludin Dingle [3], Sam F. Greenbury [4,5] Chico Q. Camargo [6] Jonathan P. K. Doye [7] Sebastian E. Ahnert [4,8,9] and Ard A. Louis [1*]

(Dated: July 27, 2021)

Engineers routinely design systems to be modular and symmetric in order to increase robustness to perturbations and to facilitate alterations at a later date. Biological structures also frequently exhibit modularity and symmetry, but the origin of such trends is much less well understood. It can be tempting to assume – by analogy to engineering design – that symmetry and modularity arise from natural selection. But evolution, unlike engineers, cannot plan ahead, and so these traits must also afford some immediate selective advantage which is hard to reconcile with the breadth of systems where symmetry is observed. Here we introduce an alternative non-adaptive hypothesis based on an algorithmic picture of evolution. It suggests that symmetric structures preferentially arise not just due to natural selection, but also because they require less specific information to encode, and are therefore much more likely to appear as phenotypic variation through random mutations. Arguments from algorithmic information theory can formalise this intuition, leading to the prediction that many genotype-phenotype maps are exponentially biased towards phenotypes with low descriptional complexity. A preference for symmetry is a special case of this bias towards compressible descriptions. We test these predictions with extensive biological data, showing that that protein complexes, RNA secondary structures, and a model gene-regulatory network all exhibit the expected exponential bias towards simpler (and more symmetric) phenotypes. Lower descriptional complexity also correlates with higher mutational robustness, which may aid the evolution of complex modular assemblies of multiple components.

Evolution proceeds through genetic mutations which generate the novel phenotypic variation upon which natural selection can act. The relationship between the space of genotypes and the space of phenotypes can be encapsulated as a genotype-phenotype (GP) map [1–3]. These can be viewed algorithmically, where random genetic mutations search in the space of (developmental) algorithms encoded by the GP map, a relationship that has been highlighted, for example, in plants [4], in Dawkins' 'biomorphs' [5] and in molecules [6].

Genetic mutations are random in the sense that they occur independently of the phenotypic variation they produce. This does not, however, mean that the probability $P(p)$ that a GP map produces a phenotype p upon random sampling of genotypes will be anything like a uniformly random distribution. Instead, highly general (but rather abstract) arguments based on the coding theorem of algorithmic information theory (AIT) [7], predict that the $P(p)$ of many GP maps should be highly biased towards phenotypes with low Kolmogorov complexity $K(p)$ [8]. High symmetry can, in turn, be linked to low $K(p)$ [6, 9–11]. An intuitive explanation for this algoritmic bias towards symmetry proceeds in two steps: 1.) Symmetric phenotypes typically need less information to encode algorithmically, due to repetition of subunits. This higher compressibility reduces constraints on genotypes, implying that more genotypes will map to simpler, more symmetric phenotypes than to more complex asymmetric ones [2, 3]. 2.) Upon random mutations these symmetric phenotypes are much more likely to arise as potential variation [12, 13], so that a strong bias towards symmetry may emerge even without natural selection for symmetry.

Symmetry in protein quaternary structure and polyominoes

We first explore evidence for this algorithmic hypothesis by studying protein quaternary structure, which describes the multimeric complexes into which many proteins self-assemble in order to perform key cellular functions (Fig. 1A and Supporting Information (SI) Fig S1 and section S1). These complexes can form in the cell if proteins evolve attractive interfaces allowing them to bind to each other [14–16]. We analysed a curated set of 34,287 protein complexes extracted from the Protein Data Bank (PDB) that were categorised into 120 different bonding topologies [16]. In Fig. 1B, we plot, for all complexes involving 6 subunits (6-mers), the frequency with which a protein complex of topology p appears against the descriptional complexity $\tilde{K}(p)$, an approximate measure of its true Kolmogorov assembly complexity $K(p)$, defined here as the minimal number of distinct interfaces required to assemble the given structure under general self-assembly rules (Methods). Here $\tilde{K}(p)$ can also be thought of as a measure of the minimal number of evolutionary innovations needed to make a self-assembling complex. The highest probability structures all have relatively low $\tilde{K}(p)$. Since structures with higher symmetry need less information to describe [6, 9–11], the most frequently observed

* ard.louis@physics.ox.ac.uk [1]Rudolf Peierls Centre for Theoretical Physics, University of Oxford; Oxford, UK. [2]Department of Mathematics and Computational Biology Unit, University of Bergen; Norway. [3]Centre for Applied Mathematics and Bioinformatics, Department of Mathematics and Natural Sciences, Gulf University for Science and Technology; Kuwait. [4]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge; Cambridge, UK. [5]Department of Metabolism, Digestion and Reproduction, Imperial College London; London, UK. [6]Dept of Computer Science, University of Exeter; Exeter, UK. [7]Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford; Oxford, UK. [8]Department of Chemical Engineering and Biotechnology, University of Cambridge; Cambridge, UK. [9]The Alan Turing Institute, British Library; London, UK.

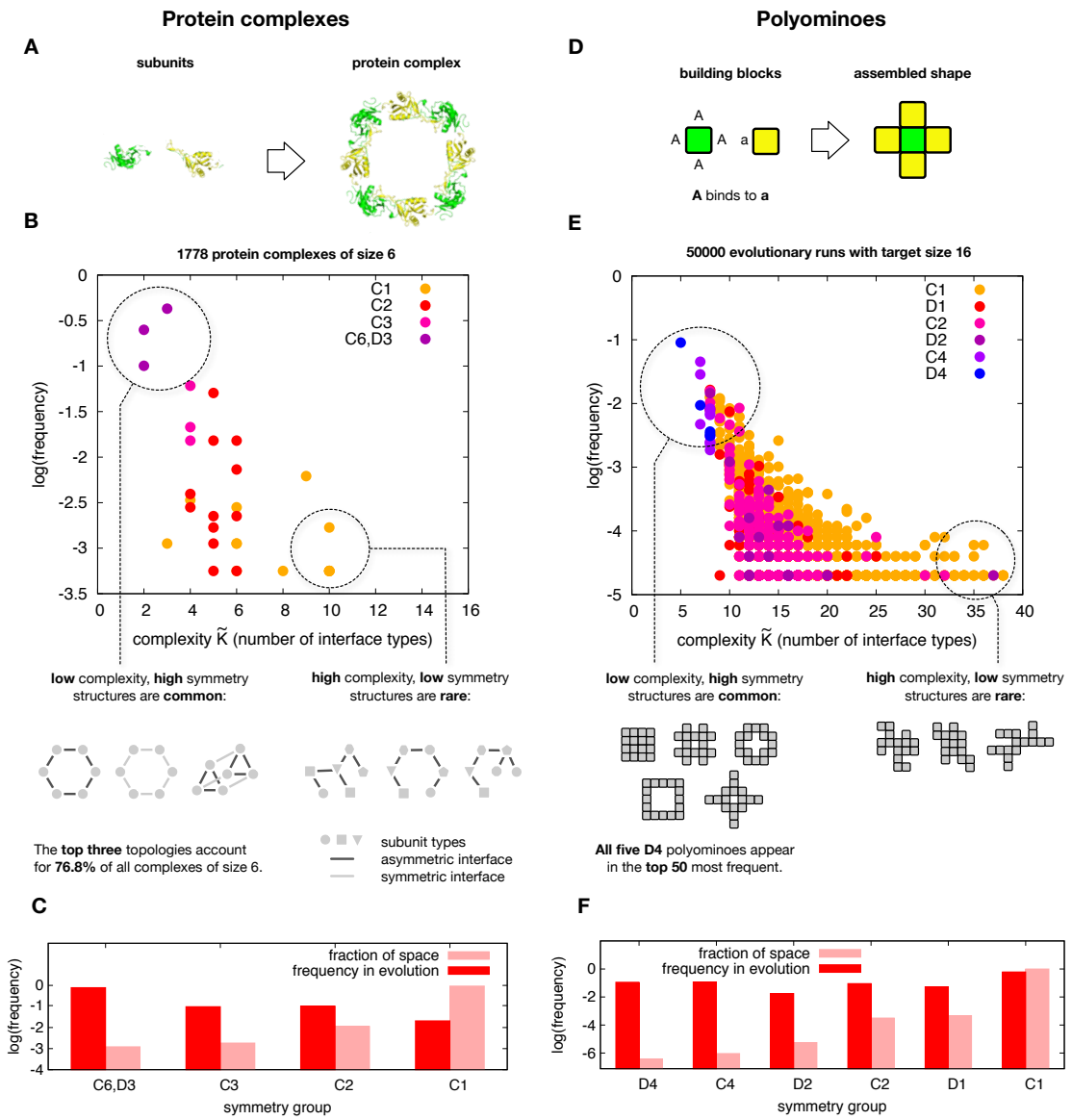

FIG. 1. (A) Protein complexes self-assemble from individual units. (B) Frequency of 6-mer protein-complex topologies found in the PDB versus the number of interface types, a measure of complexity $\tilde{K}(p)$. Symmetry groups are in standard Schoenflies notation: C_6, D_3, C_3, C_2, C_1. There is a strong preference for low complexity/high symmetry structures. (C) Histograms of the frequencies of symmetries for 6-mer topologies found in the PDB (dark red) versus the frequencies by symmetry of the morphospace of all possible 6-mers illustrate that symmetric structures are hugely over-represented in the PDB database. (D) Polyomino complexes (here a binds to A) self-assemble from individual units just as the proteins do. (E) The frequency of polyominoes that fix in evolutionary simulations with a fitness maximum at 16-mers versus the number of interface types (a measure of complexity $\tilde{K}(p)$) exhibits a strong bias towards high symmetry structures, similar to protein complexes. (F) Histograms of the frequency of symmetry groups for all 16-ominoes (light) and for 16-ominoes appearing in the evolutionary runs (dark), quantify how strongly biased variation drives a pronounced preference for high symmetry structures.

complexes are also highly symmetric. Figs. 1C and Figs S2A & S3A further demonstrate that structures found in the PDB are significantly more symmetric than the set of all possible 6-mers (Methods). Similar biases towards high symmetry structures obtain for other sizes (Fig S2B).

In order to understand the evolutionary origins of this bias towards symmetry we turn to a tractable GP map for protein quaternary structure. In the Polyomino GP map, two-dimensional tiles self-assemble into polyomino structures [17] that model protein-complex topologies [18] (Fig. 1D). The sides represent the interfaces that bind proteins together. Within the Polyomino GP map, the genomes are bit strings used to describe a set of the tiles and their interactions. The phenotypes are polyomino shapes p that emerge from the self-

assembly process. Although this model is highly simplified, it has successfully explained evolutionary trends in protein quaternary structure such as the preference of dihedral over cyclic symmetry in homomeric tetramers [15, 17], or the propensity of proteins to form larger aggregates such as haemoglobin aggregation in sickle-cell anaemia [18].

To explore the strong preference for simple structures, we performed evolutionary simulations where fitness is maximised for polyominoes made of 16 blocks (Methods). With 16 tile types and 64 interface types, the GP map denoted as $\mathcal{S}_{16,64}$ allows all 13,079,255 possible 16-mer polyomino topologies (SI Table I) to be made. Fig 1E demonstrates that evolutionary outcomes are exponentially biased towards 16-mer structures with low $\tilde{K}(p)$ (using the same complexity measure as for the proteins (Methods)), even though every 16-mer has the same fitness.

The extraordinary strength of the bias towards high symmetry can be further illustrated by examining the prevalence of the two highest symmetry groups in the outcomes of evolutionary simulations. For 16-mers, there are 5 possible structures in class D_4 (all symmetries of the square) and 12 in C_4 (4-fold rotational symmetry). Even though these 17 structures represent just over a millionth of all 16-mer phenotypes, they make up about 30% of the structures that fix in the evolutionary runs, demonstrating an extraordinarily strong preference for high symmetry (See also Fig S3B). Comparing the histograms in Fig. 1C and Fig. 1F shows that the polyominoes exhibit a qualitatively similar bias towards high symmetry as seen for the proteins. We checked that this strong bias towards high symmetry/low $\tilde{K}(p)$ holds for a range of other evolutionary parameters (such as mutation rate) and for other polyomino sizes, see Fig S6 and SI section S3C. Natural selection explains why 16-mers are selected for (as opposed to other sizes). But, since every 16-mer is equally fit, natural selection does not explain the remarkable preference for symmetry observed here, which is instead caused by bias in the arrival of variation.

Evolutionary simulations compared to sampling

In order to further understand the mechanisms that deliver the evolutionary preference for high symmetry, we calculated the probability $P(p)$ of obtaining phenotype (polyomino shape) p by uniformly sampling 10^8 genomes for the $\mathcal{S}_{16,64}$ GP map, and counting each time a particular structure p (which can be any size) appears. Fig. 2 shows that $P(p)$ varies over many orders of magnitude for different p. High $P(p)$ only occurs for low $\tilde{K}(p)$ structures while high $\tilde{K}(p)$ structures have low $P(p)$. The inset of Fig. 2 shows that the $P(p)$ from an evolutionary run from Fig. 1 closely follows the $P(p)$ for 16-mers from random sampling. We tested this correlation for a range of different evolutionary parameters, and also for both randomly assigned and fixed fitness functions, and always observe relationships between $P(p)$ and $\tilde{K}(p)$ that are strikingly similar to those found for random sampling (Fig. S6).

The observed similarity in all these different evolutionary regimes is predicted by the *arrival of the frequent* population dynamics framework of ref [12] (SI section S2). For highly biased GP maps, it predicts that, for a wide range of mutation rates and population sizes, the rate at which variation (phenotype p) arises in an evolving population is, to first order, directly proportional to the probability $P(p)$ of it appearing upon uniform random sampling over genotypes. Strong bias in the arrival of variation can overcome fitness differences, and so control evolutionary outcomes [12, 19]. Interestingly, recent results for deep learning support this evolution dynamics picture. Deep neural nets show a strong Occam's razor like bias towards simple outputs [20] upon random sampling of parameters, and these frequent (and simple) outputs appear with similar probability under training with stochastic gradient descent [21]. This similarity between random sampling and the outcome of a stochastic optimiser strengthens the case for extending the applicability of the arrival of the frequent framework for highly biased to maps to a wide range of fitness landscapes (see SI section S2 for fuller discussion).

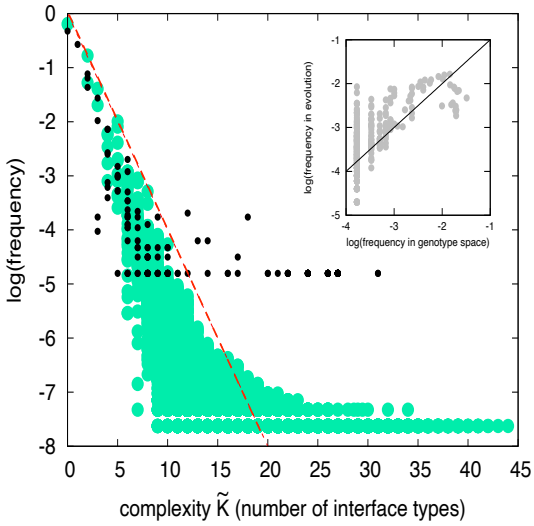

FIG. 2. The frequency that a particular protein quaternary structure topology p (black circles) appears in the PDB versus complexity $\tilde{K}(p)$=number of interface types, closely resembles the frequency distribution of all possible polyomino structures, obtained by randomly sampling 10^8 genotypes for the $\mathcal{S}_{16,64}$ space (green cirles). Simpler (more compressible) phenotypes are much more likely to occur. An illustrative AIT upper bound from Eq. (1) is shown with $a = 0.75, b = 0$ (dashed red line). *Inset*. The frequency with which particular 16-mers are found to fix in evolutionary runs from Fig. 1E is predicted by the frequency with which they arise on random sampling of genotypes; the solid line denotes $x = y$.

Fig. 2 also illustrates a striking similarity between the probability/complexity scaling for polyominoes and that of protein complex structures. Note that finite sampling effects lead to a widening of the lowest frequency outputs [8] (see also Fig. S5) suggesting that as more structures are deposited in the 3DComplex database [14] the agreement with the polyomino distribution may improve further. Given the simplicity of the polyomino model, this near quantitative agreement is probably somewhat fortuitous. Nevertheless, the arrival of the frequent mechanism, which for polyominoes explains the remarkably close similarity of the $P(p)$ v.s. $\tilde{K}(p)$ relationships across different evolutionary scenarios (see e.g. Figs S4-S9). predicts that the probability-complexity relationships for the

protein complexes will be robust, *on average*, to the many different evolutionary histories that generated these complexes. Taken together, the data and arguments above strongly favour our hypothesis that bias in the arrival of variation, and not some as yet undiscovered adaptive process, is the first order explanation of the prevalence of high symmetries in protein complexes.

Algorithmic information theory and GP maps

These results beg another question: Is the bias towards simplicity (low $\tilde{K}(p)$) observed for protein clusters and polyominoes a more general property of GP maps? Some intuition can be gleaned from the famous trope of monkeys typing at random on typewriters. If each typewriter has M keys, then every output of length N has equal probability $1/M^N$. By contrast, if the monkeys' keyboards are connected to a computer programming language then, for example, accidentally hitting the 21 characters of the program *print "01" 500 times;* will generate the $N = 1000$ digit string $010101\ldots$ with probability $1/M^{21}$ instead of $1/M^{1000}$. In other words, when searching in the space of algorithms, outputs that can be generated by short programs are exponentially more likely to be produced than outputs that can only generated by long programs.

This intuition that simpler outputs are more likely to appear upon random inputs into a computer programming language can be precisely quantified in the field of AIT [7], where the Kolmogorov complexity $K(p)$ of a string p is formally defined as the shortest program that generates p on a suitably chosen universal Turing machine (UTM). While GP maps are typically not UTMs, and strictly speaking Kolmogorov complexity is uncomputable, a relationship between the probability $P(p)$ and a computable descriptional complexity $\tilde{K}(p)$ (typically based on compression) which approximates the true $K(p)$ has recently been derived [8] for (non-UTM) input-output maps $f : I \to O$ between N_I inputs and N_O outputs. For a fairly general set of conditions, including that $N_I \gg N_O$, and that the maps are asymptotically simple (see SI section S5), the probability $P(p)$ that a map f generates output p upon random inputs can be bounded as:

$$P(p) \leq 2^{-a\tilde{K}(p)-b} \tag{1}$$

where $\tilde{K}(p)$ is an appropriate approximation to the true Kolmogorov complexity $K(p)$, and a and b are constants that depend on the map, but not on p. While Eq. (1) is only an upper bound, it can be shown [22] that outputs generated by uniform random sampling of inputs are likely to be close to the bound. In extensive tests, Eq. (1) provided accurate bounds on the $P(p)$ for systems ranging from from coupled differential equations to the RNA SS GP map [8] to deep neural networks [20], suggesting widespread applicability.

Since the number of genotypes is typically much greater than the number of phenotypes [1–3], and their relationship is encoded in a set of biophysical rules that typically depend weakly on system size, many GP maps satisfy the conditions [8] for Eq. (1) to apply (see also SI section S5). In Fig. 2, we we show an example of how Eq. (1) can act as an upper bound to $P(p)$ for the polyominoes and the protein complexes. In SI section S5C, we demonstrate that this AIT formalism also works well for other choices of the complexity $\tilde{K}(p)$, so that our results do not depend on the particular choices we make here. The AIT formalism also suggests that related systems should have similar probability-complexity relationships, which helps explain why the polyominoes and proteins have similar $P(p)$ v.s. $\tilde{K}(p)$ plots.

Since many GP maps satisfy the conditions for simplicity bias, including those where symmetry may be harder to define, we therefore hypothesised that a bias towards simplicity may also strongly affect evolutionary outcomes for many other GP maps in nature. We tested this hypothesis for RNA secondary structure and a model GRN.

Simplicity bias in RNA secondary structure

Because it can fold into well-defined structures, RNA is a versatile molecule that performs many biologically functional roles besides encoding information. While sequence to 3D structure prediction is hard to solve computationally, a simpler problem of predicting secondary structure (SS), which describes the bonding pattern of the bases, can be both accurately and efficiently calculated [24]. The map from sequences to SS is perhaps the best-studied GP map, and has provided many conceptual insights into the role of structured variation in evolution [1–3, 12, 25–27]. It has already been shown, see e.g. [26–28], that the highly biased RNA GP map strongly determines the distributions of RNA shape properties in the fRNAdb database [23] of naturally occurring non-coding RNA (ncRNA). Although natural selection still plays a role (see [26, 27] for further discussions), the dominant determinant of these structural properties is strong bias in the arrival of variation [12]. It was recently shown [8] that the RNA SS GP map is well described by Eq. (1). Combining these observations leads to the hypothesis that functional ncRNA in nature should also be exponentially biased towards more compressible low $\tilde{K}(p)$ structures.

To test this hypothesis, we first, for length $L = 30$, calculate $\tilde{K}(p)$ with a standard Lempel-Ziv compression technique [8] to directly measure the descriptional complexity of the dot-bracket notation of a SS (Methods and SI section S4). Fig. 3A shows that there is a strong inverse correlation between frequency and complexity for both naturally occurring and randomly sampled phenotypes (note that $L = 30$ is quite short so that finite size effects are expected [8] to affect the correlation with Eq. (1)). For longer RNA, the agreement with Eq. (1) is better (see e.g. ref [8], and Figs. S10, and S11. For $L = 30$ there are about 3×10^6 possible SS [26], but only 17,603 are found in the fRNAdb database [23], and these are much more likely to be more compressible low $\tilde{K}(p)$ structures. Fig. 3C shows that randomly sampling of sequences provides a good predictor for the frequency with which these structures are found in the database, consistent with previous observations [26, 27] and the arrival of the frequent framework [12].

For lengths longer than $L = 30$, the databases of natural RNAs show little to no repeated SS, so individual frequencies cannot be extracted. To make progress, we apply a well established coarse-graining strategy that recursively groups together RNA structures by basic properties of their shapes [29], which was applied to naturally occurring RNA SS in ref. [27].

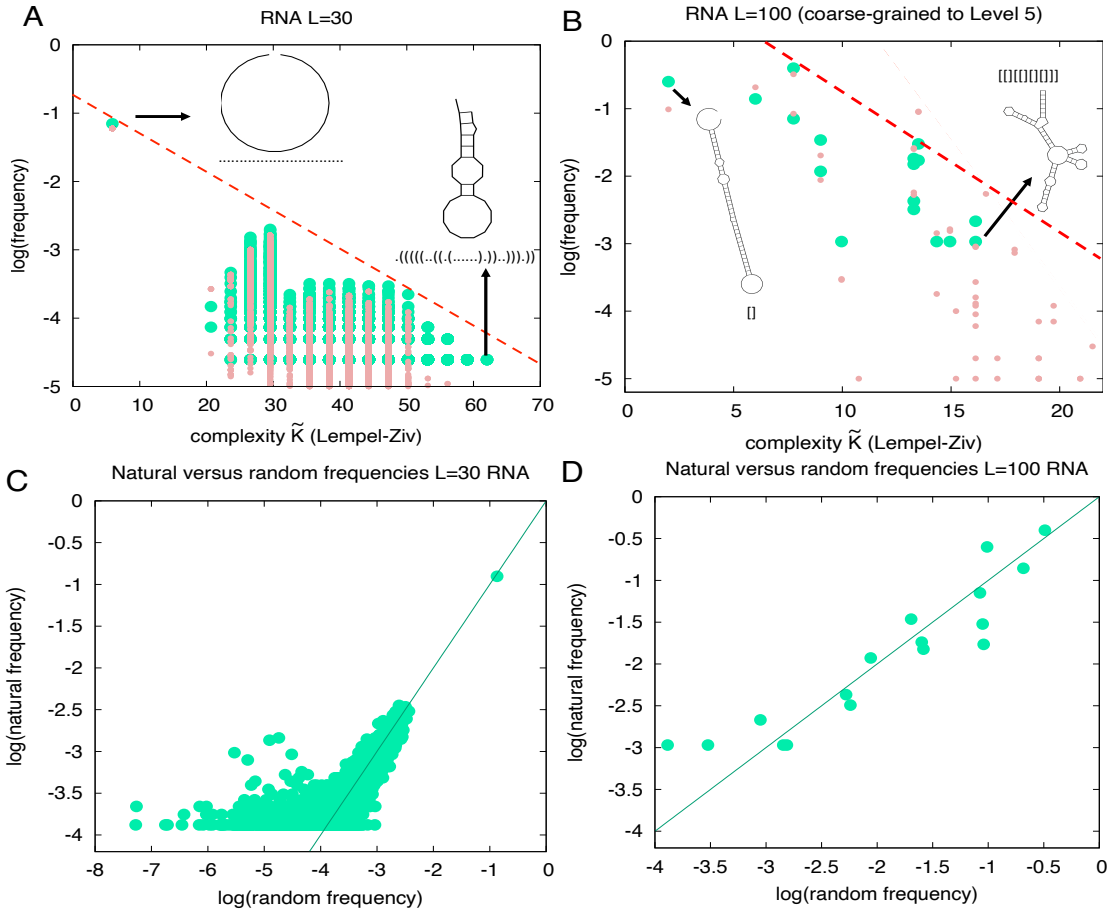

FIG. 3. Frequency/probability versus complexity $\tilde{K}(p)$ for (A) $L = 30$ RNA full SS and (B) $L = 100$ SS coarse-grained to level 5 (Methods). Probabilities for structures taken from random sampling of sequences (light red) compare well to the frequency found in the fRNA database [23] (green dots) for 40,554 functional $L = 30$ RNA sequences with 17,603 unique dot-bracket SS and for 932 natural $L = 100$ RNA sequences mapping to 16 unique coarse-grained level 5 structures. The dashed lines show a possible upper bound from Eq. (1). Examples of high probability/low complexity and low probability/high complexity SS are also shown. In (C) and (D) we directly compare the frequency of RNA structures in the fRNAdb database to the frequency of structures upon uniform random sampling of genotypes for $L = 30$ SS and $L = 100$ coarse-grained structures respectively. The lines are $y = x$. Correlation coefficients are 0.71 and 0.92, for L=30 and L=100 respectively, with p-value$< 10^{-6}$ for both. Sampling errors are larger at low frequencies.

At the highest level of coarse-graining (level 5) there are many repeat structures in the fRNAdb database, allowing for frequencies to be directly measured (Methods). For $L = 100$ we compare the empirical frequencies to $P(p)$ estimated by random sampling. Fig. 3B shows that there is again a strong negative correlation between frequency and complexity. (see also SI Tables II and III and Figs. S10 & S11 for fRNAdb and Rfam database data). Fig. 3D shows that natural frequencies are well predicted by the random sampling, as seen in ref. [27] for other lengths. Again, only a tiny fraction ($\approx 1/10^8$) of all possible phenotypes is explored by nature [27]. The RNA SS GP map exhibits simplicity bias phenomenology similar to the protein complexes and the polyomino GP map. While the simpler group-theory based symmetries discussed for protein complexes and polyominoes do not apply here, the bias towards lower $\tilde{K}(p)$ reflects the more generalised symmetries in the RNA SS structures.

Model gene regulatory network

The protein and RNA phenotypes both describe shapes. Can a similar strong preference for simplicity be found for other classes of phenotypes? To answer this question, we also studied a celebrated model for the budding yeast cell-cycle [30], where the interactions between the biomolecules that regulate the cell-cycle are modelled by 60 coupled ordinary differential equations (ODEs) As a proxy for the genotypes, we randomly sample the 156 biochemical parameters of the ODEs (Methods). For each set of parameters, we calculate the complexity of the concentration versus time curve of the CLB2/SIC1 complex (a key part of the cycle) using the up-down method [31]. Fig. 4 shows that $P(p)$ exhibits an exponential bias towards low complexity time curves, as hypothesised. Of course many of these phenotypes may not supply the biological function needed for the budding yeast cell-cycle. But interestingly, the wild-type phenotype has the lowest complexity of all the phenotypes we found, and is also the most likely to arise by random mutations. While the evolutionary

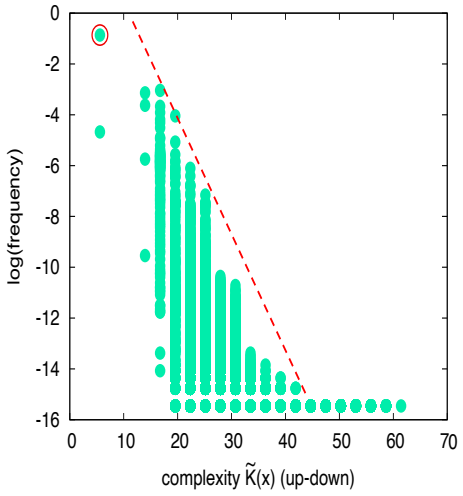

FIG. 4. $P(p)$ v.s. complexity $\tilde{K}(p)$ for the budding yeast ODE cell cycle model [30]. Phenotypes are grouped by complexity of the time-output of the key CLB2/SIC1 complex concentration. Higher $P(p)$ means a larger fraction of parameters generate this time-curve. The red circle denotes the wild-type phenotype, which is one of the simplest and most likely phenotypes to appear. The dashed show a possible upper bound from Eq. (1). There is a clear bias towards low complexity outputs.

origins of this GRN are complex, we again suggest that a bias towards simplicity in the arrival of variation played a key role in its emergence.

Discussion

Our two main hypotheses are: 1.) GP maps are, under random mutations, exponentially biased towards phenotypic variation with low descriptional complexity, as predicted by AIT [8]. 2.) Such strong bias in the arrival of variation can affect adaptive evolutionary dynamics, leading to a much higher prevalence of low complexity (high symmetry) phenotypes than can be explained by natural selection alone.

The arguments above are general enough to suggest that many biological systems, beyond the examples we provided, may favour simplicity and, where relevant, high symmetry, without requiring selective advantages for these features. For example, there are claims that HP lattice proteins with larger $P(p)$ are typically more symmetric [32], and similar patterns have been suggested for protein tertiary structure in the PDB [33]. In SI section S6 we present further evidence that protein tertiary structure, signalling networks [34] and Boolean threshold models for GRNs [35] also exhibit bias in the arrival of variation. At a more macroscopic level, a model of tooth development [36] suggests that simpler phenotypes evolved earlier, consistent with a high encounter probability in evolutionary search. Similarly, for both teeth [37] and leaf shape [38], mutations to more simpler tooth phenotypes are more likely than mutations to more complex phenotypes, an effect our theory also predicts. A recent theoretical study [39] of the development of morphology, which also found that simple morphologies were more likely to appear than complex ones upon random parameter choices. The L-systems used to model plant development [4] show simplity bias [8], and Azevedo *et al.* [40] showed that developmental pathways for cell lineages are significantly simpler (in a Kolmogorov complexity sense) than would be expected by chance.

On the other hand, for complex phenotypic traits affected by many loci, variation may be more isotropic so that bias is weak. For such traits, where classical population genetics – which focuses on shifting allele frequencies in a gene pool where standing variation is abundant – typically works well, our arguments may no longer hold. The phenotype bias we discuss here is fundamentally about the origin of novel variation [19, 41], and so is most relevant on longer time-scales.

Finally, simple systems have a larger $P(p)$ and are therefore more mutationally robust [1–3, 26, 42] (see also SI section S4B). A correlation between low complexity and robustness is also found in the engineering literature [42, 43]. Biological complexity often arises from connecting existing components together into modular wholes. If the individual components are more robust, then it is easier for them to evolve additional function, for example a patch to bind to another protein, without compromising their core function. Similarly, a larger robustness may also enhance the ability of a system to encode cryptic variation, facilitating access to new phenotypes [44]. A natural tendency towards simpler and more robust structures may therefore facilitate the emergence of modularity, where individual components can evolve independently [45], and so make living systems more globally evolvable.

METHODS

Protein-complex topologies

Our analysis of protein quaternary structure builds upon the techniques and data presented in ref. [16], where a curated set of 30,469 monomers, 28,860 homomers, and 5,527 heteromers were extracted from the Protein Data Bank (PDB), and classified into 120 distinct topologies. These were then used to make a periodic table of possible topologies. Protein complexes are described in terms of a weighted subunit interaction graph. An illustration of the topologies, and how they are generated is shown in Fig S1, for two heteromeric complexes, and their final graph topologies. Further examples of topologies and the PDB structures they describe can be found at http://www.periodicproteincomplexes.org/. The nodes of the graph are labelled according to their protein identities and the weights of the connections are the interface sizes in Å^2. The procedure for enumerating possible topologies, and for classifying existing and potential topologies is described more fully in ref. [16]. This approach only considers the largest interfaces, which if cut would disconnect the complex. The reason is that small interfaces that can be cut without disconnecting the complex are likely to be circumstantial, and unlikely to play an important role in the assembly and evolution of the complex. After constructing the weighted subunit interaction graphs in this manner we identify the topologically distinct interaction graph of subunit types (see for example Fig. S1C, with the additional distinction between symmetric and asymmetric self-interactions of a subunit type, corresponding to homomeric interfaces.

We take the number of interface types of protein complex p to be the complexity measure $\tilde{K}(p)$. This choice is proportional to the number of individual mutations needed to generate the self-assembled complex. See SI section S5C for a longer discussion of different possible complexity measures. Unlike the polyomino case, where the building block is a square tile, the geometry of an individual protein is highly variable. For example, a cyclic homomeric 6-ring and a cyclic homomeric 10-ring will have the same topologically distinct interface configuration (which is just the two parts of the same asymmetric interface on a single subunit). This will be distinct from a heteromeric 6-ring in which we have two halves of two different symmetric interfaces on a subunit, and also distinct from a simple heterodimer. All three

of these however have the same number of interface types (2) and so appear at $\tilde{K}(p) = 2$ in the distribution of Figs. 1 and 2 in the main text. The single point that appears at $\tilde{K}(p) = 1$ for Fig. 2 is a homodimer and the single point at $\tilde{K}(p) = 0$ is a monomer. The symmetries of all protein complexes presented here are taken directly from the PDB.

To calculate the symmetries of all hypothetical protein complexes of size six in Fig. 1C, we used the following procedure. We first consider all topologically distinct graphs of size six with up to six different subunit types and symmetric or asymmetric homomeric interfaces between subunits of the same type. By comparing all 6! possible permutations of the adjacency matrix and the associated node labels we then calculate the permutation symmetries of the node types on these graphs as a proxy for the spatial symmetry of the hypothetical protein complexes that they represent. This collapses D3 and C6 into one category, but allows us to distinguish this category from C3, C2, and C1 (and these from each other). Further discussion of the protein complexes can be found in SI section S1.

Polyominoes

The polyomino model was implemented as described in refs. [6, 17, 18]. The genome encodes a ruleset consisting of $4n$ numbers which describe the interactions on each edge of n square tiles. Each number is represented as a length b binary string, so that the whole genome is a binary string of length $L = 4nb$. The interactions bond irreversibly and with equal strength in unique pairs ($1 \leftrightarrow 2$, $3 \leftrightarrow 4$, ...), with types 0 and $2^b - 1$ being neutral, not bonding to any other types. We label a given polyomino GP map with up to n possible tiles and $4n$ possible colors as $\mathcal{S}_{n,4n}$; in this paper we usually work with $\mathcal{S}_{16,64}$.

The assembly process is initiated by placing a single copy of the first-encoded subunit tile on an infinite grid. A different protocol where any tile may be used to seed the assembly is also possible and does not significantly affect the results presented here. Assembly then proceeds as follows. 1) Available moves are identified, consisting of an empty grid site, a particular tile and a particular orientation, such that placing that tile in that orientation in the site will form a bond to an adjacent tile that has already been placed. 2) If there are no available moves, terminate assembly. 3) Choose a random available move and place the given tile in that orientation at that site. 4) If the current structure has exceeded a given cutoff size, terminate assembly. 5) Go to step 1.

This process is repeated 20 times to ensure that assembly is *deterministic* – that is, that the same structure is produced each time. If different structures are produced, or the structure exceeds a cutoff size (here taken to be larger than a 16×16 grid), the structure is placed in the category 'UND' (unbounded or non-deterministic). For the calculation of probabilities/frequencies $P(p)$ we ignore genotypes that produce the UND phenotype. This choice mimics the intuition that unbounded protein assemblies, or else proteins that do not robustly self-assemble into the same shape, are highly deleterious.

The ruleset $\mathcal{S}_{16,64}$ allows any 16-mer to be made, since it is always possible to use addressable assembly where each tile is unique to a specific location. But many 16-mers can be made with significantly fewer than 16 tile types, although there are examples that (to our knowledge) can only be made with all 16 tiles, so that a space allowing up to 16 tiles is needed.

To assign complexity values for the polyominoes, a measure similar to that used for the proteins was applied. First, the minimal complexity over the different genomes that generate polyomino p is estimated by a sampling and finding the shortest rule set and removing redundant information. The search for a minimal complexity genome will be more accurate for high probability polyominoes than for low probability polyominoes. We checked that for most structures only a fairly limited amount of sampling provided an accurate estimate of the minimal complexity; the minimal complexity genome is typically the most likely to be found. The effects of finite sampling are illustrated further in SI section S3 and Fig. S5. The complexity $\tilde{K}(p)$ of polyomino p is then given by the smallest number of unique edge labels (interface types) in the minimal genomes – thus, twice the number of hetero-interactions, just as in the protein system above. A longer discussion of different choices of complexity measure, showing that the qualitative behaviour is not very sensitive to details in the choice of approximate measure of algorithmic (Kolmogorov) information, can be found in SI Section S3C.

Evolutionary simulations of polyomino structures are performed following methods described in ref. [17]: A population of N binary polyomino genomes is maintained at each time-step. The assembly process is performed for each genome and the resulting structure is recorded. UND genomes are assigned zero fitness. Other structures are assigned a fitness value based on the applied fitness function. These fitness values are used to perform roulette wheel selection, whereby a genome g_i with fitness $f(g_i)$ is selected with probability $f(g_i)/\sum_j f(g_j)$. Selection is performed N times (with replacement) to build the population for the next time-step. Selected genomes are cloned to the next generation, then point mutations are applied with probability μ at each locus. A point mutation changes a 0 to a 1 and vice versa in the genome. We do not employ crossover or elitism in these simulations.

We employ several different fitness functions. In the *unit fitness* protocol, all polyomino structures that are not UND are assigned fitness 1. In the *random fitness* protocol, each polyomino structure is assigned a fitness value uniformly randomly distributed on $[0, 1]$, and these values are reassigned for each individual evolutionary run. In the *size fitness* protocol, a polyomino of size s has fitness $1/(|s - s^*| + 1)$, so that polyominoes of size s^* have unit fitness and other sizes have fitness decreasing with distance from s^*. The simulations for Fig. 1E were done with $N = 100$ and $\mu = 0.1$ per genome, per generation. A number of other evolutionary parameters are compared in SI section 3C and Fig. S6, showing that our main result – that the outcome of evolutionary dynamics exhibit an exponential bias towards simple structures – is not very sensitive to details such as mutation rate or the choice of fitness function.

RNA secondary structure GP map

For $L = 30$ RNA we randomly generated 32,000 sequences, and for $L = 100$, we generated 100,000 random sequences. As in refs. [26, 27], secondary structure (SS) is computationally predicted using the `fold` routine of the Vienna package [24] based on standard thermodynamics of folding. All folding was performed with parameters set to their default values (in particular, the temperature is set at $T = 37°$C). We then calculated the neutral set size (NSS(p)), the number of sequences mapping to a SS p, for each SS found by random sampling, by using the neutral network size estimator (NNSE) described in ref [28], which is known to be quite accurate for larger NSS structures [26]. We used default settings except for the total number of measurements (set with the -m option) which we set to 1 instead of the default 10, for the sake of speed, but this does not noticeably affect the outcomes we present here.

RNA structures can be represented in standard dot-bracket notation, where brackets denote bonds, and dots denote unbonded pairs. For example, ...((...)).... means that the first three bases are not bonded, the fourth and fifth are bonded, the sixth through ninth are unboded, the tenth base is bonded to the fifth base, the eleventh base is bonded to the fourth base, and the final four bases are unbounded. For shorter strands such as $L = 30$, the same SS can be found multiple times in the fRNAdb.

For longer strands, finding multiple examples of the same SS becomes more rare, so that SS frequencies cannot be easily directly extracted from the fRNAdb. However, it seems reasonable, especially for larger structures, that fine details of the structures are not as important as certain more gross structural features that are captured by a more coarse grained picture of the structure. In this spirit, we make use of the well known RNA abstract shape method [29] where the dot-bracket SS are abstracted to one of five hierarchical levels, of increasing abstraction, by ignoring details such as the length of loops, but including broad shape features. For the $L = 100$ data we choose the fifth, or highest level of abstraction which only measures the stem arrangement. This choice of level is needed to achieve multiple examples of the same structure in the fRNAdb database, so that a frequency can be directly determined with statistical significance. The SS were converted to abstract shapes with the online tool available at `https://bibiserv.cebitec.uni-bielefeld.de/rnashapes`. Using these coarse-grained structures means that the theoretical probability $P(p)$ can be directly calculated from random sampling of sequences, where N_G is the number of sequences, which for an RNA GP map for length L RNA is given by $N_G = 4^L$. A similar calculation of the $P(p)$ for RNA structures for L from 40 to 126 at different levels of coarse-graining can be found in [27].

To generate the distributions of natural RNA we took all available sequences of $L = 30$ and $L = 100$ from the non-coding functional RNA database (fRNAdb [23]). As in ref. [26], we removed a small fraction ($\sim$1%) of the natural RNA sequences containing non-standard nucleotide letters, e.g. 'N' or 'R' because the standard folding packages cannot treat them. Similarly, a small fraction ($\sim$2%) of sequences were also discarded due to the neutral set size estimator (NSSE) failing to calculate the NSS (this is only

relevant for $L = 30$). We have further checked that removing by hand any sequences that were assigned putative roles, or are clear repeats, does not significantly affect the strong correlation between the frequencies found in the fRNA database and those obtained upon random sampling of genotypes. For a further discussion of the question of how well frequency in the databases tracks the frequency in nature, see also refs. [26, 27] and Fig. S10 where a comparison with the Rfam database is also made. Note that the similar behaviour we find across structure prediction methods, strand lengths, and databases would be extremely odd if artificial biases were strong on average in the fRNA database. We used $40,554$ unique RNA sequences of $L = 30$, taken from the fRNAdb, corresponding to 17,603 unique dot-bracket structures. Similarly, we used 932 unique fRNAdb $L = 100$ RNA sequences, corresponding to 17 unique level 5 abstract structures/shapes.

To estimate the complexity of an RNA SS, we first converted the dot-bracket representation of the structure into a binary string p, and then used the Lempel-Ziv based complexity measure from ref. [8] to estimate its complexity. To convert to binary strings, we replaced each dot with the bits 00, each left-bracket with the bits 10, and each right-bracket with 01. Thus an RNA SS of length n becomes a bit-string of length $2n$. Because level 5 abstraction only contains left and right brackets, i.e. [and], we simply convert left-bracket to 0, and right to 1 before estimating the complexity of the resulting bit string via the Lempel-Ziv based complexity measure from ref. [8]. The level 5 abstract trivial shape with no bonds is written as underscore, and this we simply represented as a single 0 bit. SI section S4 provides more background on RNA structures, and section S5B more detail of the complexity measure.

GRN of budding yeast cell-cycle

The budding yeast (*S. cerevisiae*) cell-cycle GRN system from ref. [30] consists of 60 coupled ordinary differential equations (ODEs) relating 156 biochemical parameters. The model parameter space (i.e. genotype space) was sampled by picking random values for each of the parameters by multiplying the wild-type value by one of $\{0.25, 0.50, \ldots, 1.75, 2.00\}$, chosen with uniform probability. The ODEs generate concentration-time curves for different biochemicals involved in cell-cycle regulations. All runs were first simulated for 1000 time steps, with every time step corresponding to 1 minute. Next, we identified the period of every run (usually on the order of 90 time steps), took one full oscillation and coarse-grained it to 50 time steps. This way, if two genotypes produce curves which are identical up to changes in period, they should ultimately produce identical or nearly-identical time series and binary string phenotypes. For every "genotype" or set of parameters, the curves for the CLB2/SIC1 complex are then discretised into binary strings using the "up-down" method [31]: for every discrete value of $t = \delta t, 2\delta t, 3\delta t, \ldots$, we calculate the slope dy/dt of the concentration curve, and if $dy/dt \geq 0$, a 1 gets assigned to the j-th bit of the output string, otherwise, a 0 is assigned to it. All strings with the same up/down profile were classified as one phenotype. To generate the $P(p)$ in Fig. 4, 5×10^6 inputs were sampled. Complexity $\tilde{K}(p)$ is assigned by using the Lempel Ziv measure from ref. [8] (see also SI section S5B) applied to binary output strings. As shown in ref. [8], this methodology works well for coupled differential equations, and the choice of input discretisation, sample size and initial conditions does not qualitatively affect the probability-complexity relationships obtained. The wildtype curve can be observed in in Fig. 2 of ref. [30] where it is labelled $\mathrm{Clb2}_T$.

Author Contributions

IGJ and SG did the polyomino simulations, SEA generated the protein data, KD generated the RNA data, KD and CQC analysed the GRN, JPKD, SEA and AAL supervised the polyomino work, AAL supervised the RNA and GRN work. All authors helped analyse the data and write the paper.

Acknowledgements

We thank N. Martin, V. Mohanty, J. Bohlin, S. Schaper and H. Zenil for discussions.

[1] Andreas Wagner. *Arrival of the Fittest: Solving Evolution's Greatest Puzzle*. Penguin, 2014.

[2] Sebastian Edmund Ahnert. Structural properties of genotype–phenotype maps. *Journal of The Royal Society Interface*, 14(132):20170275, 2017.

[3] Susanna Manrubia, José A Cuesta, Jacobo Aguirre, Sebastian E Ahnert, Lee Altenberg, Alejandro V Cano, Pablo Catalán, Ramon Diaz-Uriarte, Santiago F Elena, Juan Antonio García-Martín, et al. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Physics of Life Reviews*, 37, 2021.

[4] Przemyslaw Prusinkiewicz, Yvette Erasmus, Brendan Lane, Lawrence D Harder, and Enrico Coen. Evolution and development of inflorescence architectures. *Science*, 316(5830):1452–1456, 2007.

[5] Richard Dawkins. The evolution of evolvability. In *On Growth, Form and Computers*. Elsevier, 2003.

[6] Sebastian E. Ahnert, Iain G. Johnston, Thomas M.A. Fink, Jonathan P.K. Doye, and Ard A Louis. Self-assembly, modularity, and physical complexity. *Physical Review E*, 82(2):026117, 2010.

[7] M. Li and P.M.B. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag New York Inc, 2008.

[8] Kamaludin Dingle, Chico Q Camargo, and Ard A Louis. Input–output maps are strongly biased towards simple outputs. *Nature communications*, 9(1):761, 2018.

[9] Hod Lipson. Principles of modularity, regularity, and hierarchy for scalable systems. *Journal of Biological Physics and Chemistry*, 7(4):125, 2007.

[10] Julyan HE Cartwright and Alan L Mackay. Beyond crystals: the dialectic of materials and information. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1969):2807–2822, 2012.

[11] Mark Buchanan. Instructions for assembly. *Nature Physics*, 8(8):577–577, 2012.

[12] Steffen Schaper and Ard A Louis. The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PloS One*, 9(2):e86635, 2014.

[13] 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 computational biology*, 12(3):e1004773, 2016.

[14] Emmanuel D Levy, Jose B Pereira-Leal, Cyrus Chothia, and Sarah A Teichmann. 3D complex: a structural classification of protein complexes. *PLoS computational biology*, 2(11):e155, 2006.

[15] E. D. Levy, E. B. Erba, C. V. Robinson, and S. A. Teichmann. Assembly reflects evolution of protein complexes. *Nature*, 453:1262, 2008.

[16] 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):2245, 2015.

[17] Iain G Johnston, Sebastian E Ahnert, Jonathan PK Doye, and Ard A Louis. Evolutionary dynamics in a simple model of self-assembly. *Physical Review E*, 83(6):066105, 2011.

[18] 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.

[19] L.Y. Yampolsky and A. Stoltzfus. Bias in the introduction of variation as an orienting factor in evolution. *Evolution & Development*, 3(2):73–83, 2001.

[20] Guillermo Valle-Pérez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint*, arXiv:1805.08522, 2018.

[21] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A Louis. Is SGD a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021.

[22] Kamaludin Dingle, Guillermo Valle Pérez, and Ard A Louis. Generic predictions of output probability based on complexities of inputs and outputs. *Scientific reports*, 10(1):1–9, 2020.

[23] Taishin Kin, Kouichirou Yamada, Goro Terai, Hiroaki Okida, Yasuhiko Yoshinari, Yukiteru Ono, Aya Kojima, Yuki Kimura, Takashi Komori, and Kiyoshi Asai. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, 35(suppl 1):D145–D148, 2007.

[24] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[25] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings: Biological Sciences*, 255(1344):279–284, 1994.

[26] Kamaludin Dingle, Steffen Schaper, and Ard A Louis. The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface focus*, 5(6):20150053, 2015.

[27] Kamaludin Dingle, Fatme Ghaddar, Petr Sulc, and Ard A Louis. Phenotype bias determines how RNA structures occupy the morphospace of all possible shapes. *bioRxiv preprint*, 2020.12.03.410605, 2020.

[28] T. Jorg, O.C. Martin, and A. Wagner. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC bioinformatics*, 9(1):464, 2008.

[29] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843–4851, 2004.

[30] K.C. Chen, L. Calzone, A. Csikasz-Nagy, F.R. Cross, B. Novak, and J.J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, 15(8):3841, 2004.

[31] T.M.A. Fink, K. Willbrand, and F.C.S. Brown. 1-D random landscapes and non-random data series. *EPL (Europhysics Letters)*, 79(3):38006, 2007.

[32] Tairan Wang, Jonathan Miller, Ned S Wingreen, Chao Tang, and Ken A Dill. Symmetry and designability for lattice protein models. *The Journal of Chemical Physics*, 113(18):8329–8336, 2000.

[33] J. Hartling and J. Kim. Mutational robustness and geometrical form in protein structures. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 310(3):216–226, 2008.

[34] Karthik Raman and Andreas Wagner. The evolvability of programmable hardware. *Journal of the Royal Society Interface*, 8(55):269–281, 2011.

[35] Chico Q Camargo and Ard A Louis. Boolean threshold networks as models of genotype-phenotype maps. In *Complex Networks XI*, pages 143–155. Springer, 2020.

[36] I. Salazar-Ciudad and J. Jernvall. How different types of pattern formation mechanisms affect the evolution of form and development. *Evolution and Development*, 6(1):6–16, 2004.

[37] Enni Harjunmaa, Aki Kallonen, Maria Voutilainen, Keijo Hämäläinen, Marja L Mikkola, and Jukka Jernvall. On the difficulty of increasing dental complexity. *Nature*, 483(7389):324, 2012.

[38] R. Geeta, L.M. Davalos, A. Levy, L. Bohs, M. Lavin, K. Mummenhoff, N. Sinha, and M.F. Wojciechowski. Keeping it simple: flowering plants tend to retain, and revert to, simple leaves. *New Phytologist*, 193:481–93, 2012.

[39] Pascal F Hagolani, Roland Zimm, Renske Vroomans, and Isaac Salazar-Ciudad. On the evolution and development of morphological complexity: A view from gene regulatory networks. *PLoS Computational Biology*, 17(2):e1008570, 2021.

[40] R.B.R. Azevedo, R. Lohaus, V. Braun, M. Gumbel, M. Umamaheshwar, P.M. Agapow, W. Houthoofd, U. Platzer, G. Borgonie, H.P. Meinzer, et al. The simplicity of metazoan cell lineages. *Nature*, 433(7022):152–156, 2005.

[41] David M McCandlish and Arlin Stoltzfus. Modeling evolution using the probability of fixation: history and implications. *The Quarterly Review of Biology*, 89(3):225–252, 2014.

[42] Alden H Wright and Cheyenne L Laue. Evolvability and complexity properties of the digital circuit genotype-phenotype map. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 840–848, 2021.

[43] Giota Paparistodimou, Alex Duffy, Robert Ian Whitfield, Philip Knight, and Malcolm Robb. A network science-based assessment methodology for robust modular system architectures during early conceptual design. *Journal of Engineering Design*, 31(4):179–218, 2020.

[44] Jia Zheng, Joshua L Payne, and Andreas Wagner. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science*, 365(6451):347–353, 2019.

[45] G.P. Wagner, M. Pavlicev, and J.M. Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.