Non-deterministic genotype-phenotype maps of biological self-assembly

S. Tesoro[1] and S. E. Ahnert[1,2]

[1] *Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, CB3 0HE Cambridge, UK*

[2] *Sainsbury Laboratory, University of Cambridge, Bateman Street, CB2 1LR Cambridge, UK*

In recent years large-scale studies of different genotype-phenotype (GP) maps, including those of RNA secondary structure, lattice proteins, and self-assembling Polyominoes, have revealed that these maps share structural properties. Such properties include skewed distributions of genotypes per phenotype, negative correlations between genotypic evolvability and robustness, positive correlations between phenotypic evolvability and robustness, and the fact that a majority of phenotypes can be reached from any genotype in just a few mutations. Traditionally this research has focused on deterministic GP maps, meaning that a single sequence maps to a single outcome. Here, by contrast, we consider non-deterministic GP maps, in which a single sequence can map to multiple outcomes. Most GP maps already contain such sequences, but these are typically classified as a single, undesirable phenotype for the reason that biological processes typically rely on robust transformation of sequences into biological structures and functions. For the same reason, however, non-deterministic phenotypes play an important role in diseases, and a deeper understanding of non-deterministic GP maps may therefore inform the study of their evolution. By redefining deterministic and non-deterministic Polyomino self-assembly phenotypes in terms of the pattern of possible interactions rather than the final structure we are able to calculate GP map properties for the non-deterministic part of the map, and find that they match those found in deterministic maps.

INTRODUCTION

The mapping between genotype and phenotype is of fundamental importance to biological evolution. While evolutionary selection is the driving force behind phenotypic change, the genotype-phenotype (GP) map represents the numerous constraints imposed on evolutionary outcomes by the relationship between genotypic sequence and phenotypic structure and function. Over the past three decades the structural properties of genotype-phenotype maps have been studied in great detail. Examples of such GP maps include that of RNA secondary structure [1, 2], the HP model [3, 4], and the Polyomino map [5, 6]. This work has revealed a number of fundamental properties that can be observed for the distribution of phenotypes on the point-mutation network of genotypes across these different maps [6–10]. These properties include a highly skewed distribution of genotypes per phenotypes, strong correlations between the locations of genotypes of a given phenotype (resulting in high robustness), a negative correlation between genotypic robustness and evolvability, a positive correlation between phenotypic robustness and evolvability, and short paths between any pair of phenotypes in the point-mutation network of genotypes. All the GP maps studied so far assume that the relationship between sequence and structure is deterministic, meaning that a particular genotype maps to a single phenotype. Instances in which the same sequence can lead to multiple different outcomes in terms of the final structure are generally regarded as deleterious phenotypes [10]. A limited range of non-deterministic outcomes might however be desirable from a functional point of view, as proteins can for example undergo conformational change. The existence of such alternative phenotypes may also facilitate evolution [11]. Non-determinism that leads to a limited number of alternative functional phenotypes it is sometimes described as promiscuity, for example in the context of protein or enzyme interactions [12, 13], or plasticity [11, 14], although the term 'plasticity' is also sometimes reserved for significant phenotypic change as a result of a small number of mutations [13]. In the context of self-assembly the study of non-deterministic phenotypes is also of interest because uncontrolled self-assembly in the form of protein aggregation is the hallmark of a number of diseases, such as sickle-cell anaemia [15] and amyloidosis [16].

Here we study a non-deterministic self-assembly GP map by extending the Polyomino GP map to include assembly tile sets that do not yield a single final structure. Instead of using the final structure(s) to define the phenotype, we use the assembly graphs, which are the topologically distinct graphs of tile interactions. Building on this we then show that the properties measured for deterministic genotype-phenotype maps also hold in the non-deterministic realm. This finding has important potential implications for the study of non-deterministic disease phenotypes, as the properties of these genotype-phenotype maps are likely to influence the evolution of diseases. As an example of a disease phenotype that can be modelled using Polyominoes we discuss sickle-cell haemoglobin, which forms protein aggregates in red bllood cells.

EXTENDING THE POLYOMINO GP MAP

The Polyomino GP map has been studied in some detail in recent years, and a number of general GP map properties have been shown to hold for this map too [6, 10]. In the Polyomino GP map building blocks in the form of square tiles self-assemble on a two-dimensional lattice. Starting from a seed tile, attractive interactions between tiles facilitate this assembly. These interactions take the form of integers that are assigned to each face, and specified pairs integers bind to each other. Not all tile faces need to interact, and some integers can be chosen to denote such non-interacting faces.

For a given set of building blocks the assembly proceeds in discrete timesteps. During each timestep the following three steps are repeated:

1. A tile type and a tile orientation are picked randomly.

2. An interacting face is randomly selected on the structure that has been assembled up to this point. In the first assembly step this is the seed tile.

3. If the randomly selected tile in its randomly selected orientation provides the complementary interaction to that of the randomly selected face on the structure, then the new tile binds to the existing structure. If not, then it is discarded.

As every tile has four faces, the configuration of interactions on a set of n different building blocks can be encoded as a string of $4n$ integers, representing the interactions (clockwise) on the four faces of each tile type, in sequence. Typically the integers are limited to a fixed range $0, 1, 2, ..., c$ where c is odd, 0 and c are non-interacting, and 1 binds to 2, 3 binds to 4, etc. [5, 6, 17]. Bindings are binary in strength, and irreversible.

In previous work this string of integers is the genotype, and the final assembled structure is the phenotype, providing that the set of building blocks encoded by the genotype always builds the same final structure in the stochastic assembly process outlined above. A further requirement is that the structure is bound. If a building block set can result in more than one final structure, or an unbound structure, then the corresponding genotype is mapped to a single, all-encompassing 'unbound or non-deterministic' (UND) phenotype. This is much like the unfolded phenotype in RNA secondary structure or a lattice protein sequence in the HP model that lacks a unique ground state. This approach treats the entirety of unbound and non-deterministic phenotypes as a single entity, and one that is - from a biological perspective - usually regarded as undesirable. The vast complexity of biological processes typically relies on precision and reproducibility, whereas non-deterministic and unbound self-assembly are inherently unpredictable.

Yet the space of non-deterministic and unbound self-assembly processes is of importance, both because some biological sequences can map to more than one structure (as is the case for proteins with multiple configurations) and because unbound or non-deterministic self-assembly can be the result of mutated proteins that normally undergo bound deterministic self-assembly.

In the following we will refer to all unbound and non-deterministic phenotypes as 'non-deterministic', as unbound assembly, even if it is deterministic, will eventually grow into a final shape that is determined by the boundaries it encounters in its environment. Unbound assemblies are thus also ultimately non-deterministic.

In order to examine the space of non-deterministic phenotypes we offer a new definition of phenotypes in the Polyomino model, based not on the final assembled structure, but instead on the graph of interactions between tiles, or 'assembly graph'.

Figure 1 illustrates the concept of an assembly graph for deterministic and non-deterministic building block sets.

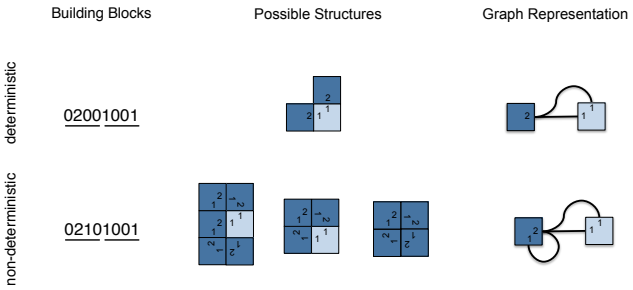

FIG. 1. Examples of deterministic (top) and non-deterministic (bottom) self-assembly phenotypes in the Polyomino GP map. A set of building blocks can be represented as a continuous string of integers (left). These building blocks self-assemble into one (in the deterministic case) or more (in the non-deterministic case) possible structures through a stochastic self-assembly process. The configurations of interactions between the building blocks can be represented as an assembly graph (right).

PHENOTYPE FREQUENCY DISTRIBUTION

Here we consider the assembly graph genotype-phenotype (AGGP) map for two tiles and eight interactions. Figure 2 shows the frequency of a phenotype versus the rank of its frequency within the map, which reveals a similar heavy-tailed distribution to those described in [6]. In contrast to the Polyomino GP map of deterministic lattice self-assembly, we are now also accounting for non-deterministic structures in our AGGP map.

The total number of topologically distinct assembly graphs with two tiles and eight interactions is 1136. Of

these, twelve yield bound deterministic structures, which make up 5,876,685 of all possible $2^{24} = 16{,}777{,}216$ degenerate genotypes, or 36.7% of all genotype space. A previous study [5] focusing on the deterministic part of this GP map classified 38.9% as bound and deterministic. The discrepancy of 2.2% between these two proportions arises (a) because the model in [5] uses a fixed seed tile as opposed to a randomly chosen one in our model, and (b) because we here define structures with internally mismatched faces (in other words structures that are only deterministic due to steric constraints, but not due to their interaction rules) as non-deterministic.

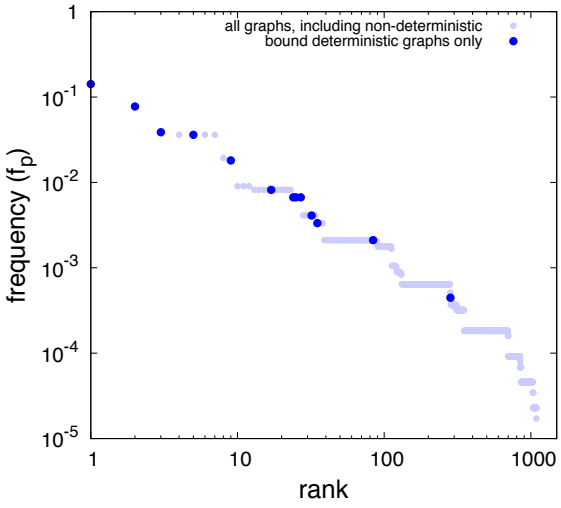

FIG. 2. Phenotypic frequency versus phenotypic frequency rank in the AGGP map for the space of two tiles and eight interactions, showing both bound deterministic (dark blue) and non-deterministic (light blue) phenotypes. The non-deterministic phenotypes display a similarly heavy-tailed distribution as the deterministic phenotypes.

GENOTYPIC AND PHENOTYPIC ROBUSTNESS

In the context of GP maps the term 'robustness' describes a resilience towards mutations that change the phenotype. [18]. Wagner defines 'genotypic robustness' ρ_g as follows [7]:

$$\rho_g = \frac{n_{p,g}}{(K-1)L} \tag{1}$$

where ρ_g is the genotypic robustness of genotype g, $n_{p,g}$ is the number of 1-mutant neighbours of g with phenotype p, K is the base of the genome (in our case the eight interactions), and L is the sequence length. There are a total of $(K-1)L$ 1-mutants for any genotype.

The robustness of the phenotype can then be defined as the average of this quantity over all genotypes with phenotype p [7]:

$$\rho_p = \frac{1}{|P|} \sum_{g \in P} \rho_g \tag{2}$$

where ρ_p is the phenotypic robustness and P is the set of genotypes with phenotype p.
Figure 3 shows that the phenotypic robustness ρ_p scales linearly with the logarithm of the phenotype frequency f_p for the AGGP map on the logarithmic scale [19], in line with previous results on the Polyomino GP map, the RNA GP map, and other GP maps [20, 21].

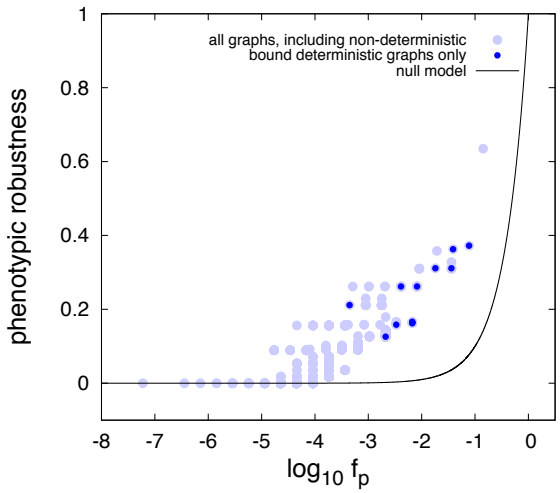

FIG. 3. Phenotypic robustness versus phenotypic frequency in the assembly graph genotype-phenotype map (AGGP) for the space of two tiles and eight interactions. Dark blue points represent deterministic phenotypes, while non-deterministic phenotypes are represented with light blue points. Both types of phenotypes follow a similar relationship of $\rho_p \propto \log f_p$, with non-deterministic phenotypes covering a larger range of frequencies. In both cases the robustness is much higher than one would expect from a random distribution of the same numbers of genotypes per phenotype [10].

EVOLVABILITY AND ROBUSTNESS

Evolvability is the ability to produce phenotypic variation [22]. At first sight, robustness and evolvability appear to be opposed, as robustness demands a lack of change in the face of mutations, whereas evolvability requires such change. Wagner [23] defined genotypic evolvability as the number of distinct phenotypes that are accessible from a given genotype through a single-point

mutation. Phenotypic evolvability can then be defined as the total number of distinct phenotypes that lie within the single-point mutation neighbourhood of a phenotype. Wagner showed that evolvability and robustness correlate positively at the phenotypic level and negatively at the genotypic level in RNA secondary structure [7]. This result has been reproduced in several other GP maps, including Polyominoes [6, 24]. As can be observed in Figures 4 and 5 this relationship also holds for the non-deterministic phenotypes of the Polyomino GP map.

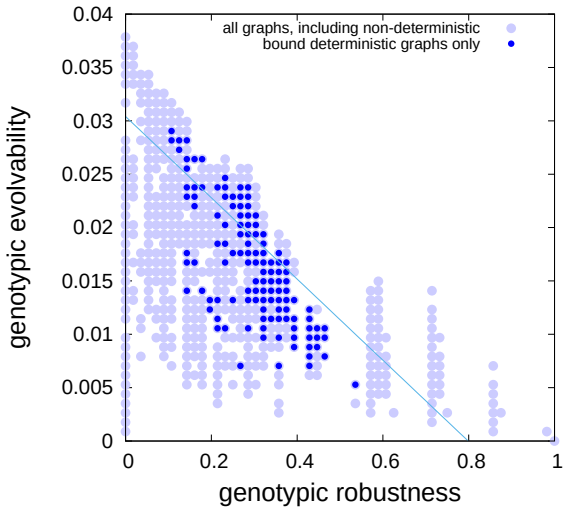

FIG. 4. Genotypic evolvability versus genotypic robustness in the AGGP map for the space of two tiles and eight interactions. The two properties are negatively correlated for both deterministic (dark blue) and non-deterministic (light blue) genotypes.

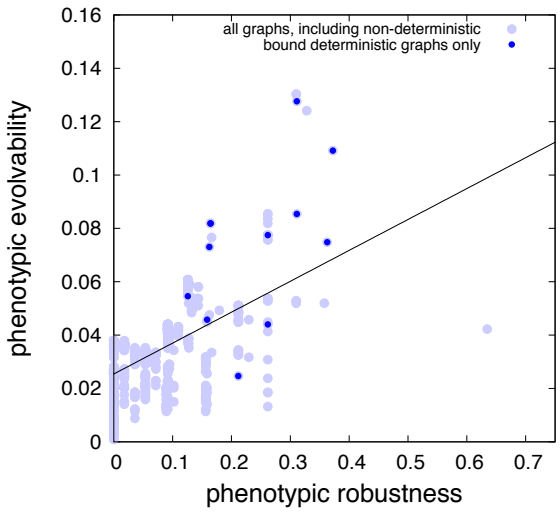

FIG. 5. Phenotypic evolvability versus phenotypic robustness in the AGGP map for the space of two tiles and eight interactions. The two properties are positively correlated for both deterministic (dark blue) and non-deterministic (light blue) phenotypes.

SHAPE SPACE COVERING

Shape space covering describes the accessibility of phenotypes in the GP map by measuring the average fraction of phenotypes that lie within n mutation steps of a randomly selected genotype [25, 26].

Figure 6 shows that the accessible fraction of phenotypes is a sigmoidal function of the number of mutations n, as observed in RNA secondary structure and the HP model of lattice proteins [26] and previous studies of the deterministic Polyomino model [6]. The majority of phenotypes can be reached within four mutation steps, starting from a randomly chosen genotype.

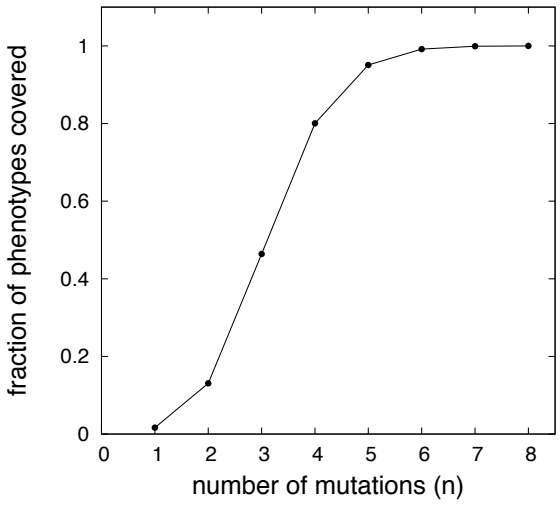

FIG. 6. Shape space covering in the AGGP map, expressed as the mean fraction of phenotypes that are accessible from a randomly selected genotype for the space of two tiles and eight interactions. On average more than half of the available phenotypes lie within four mutations of a randomly chosen genotype.

DISCUSSION AND CONCLUSION

We show here that several well-established structural properties of deterministic GP maps also hold for the non-deterministic phenotypes of the Polyomino GP map. The phenotypes are defined in terms of the configurations of interactions between the Polyomino building blocks. Recent work has shown that the structural properties of GP maps are likely to arise from the organisation of biological information into constrained and unconstrained sequences, paired with non-local effects of mutations on these levels of constraint [27]. The fact that these properties are observed for the non-deterministic phenotypes of the Polyomino GP map confirms that phenotypes defined as configurations of interactions, whether they are deterministic or non-deterministic, will exhibit the same universal structural characteristics in the GP map.

An alternative approach for the treatment of non-deterministic phenotypes is to associate multiple phenotypes with the single non-deterministic genotype from which they arise, in form of a many-to-many genotype phenotype map. Structural properties of such maps have begun to be considered in the recent literature [11], particularly in the context of plasticity, which has been also studied in genotype-phenotype maps more generally [12–14, 28–30]. Much of this work focuses on examples in which a genotype can yield a limited number of different functional phenotypes, such as enzymes that act as catalysts in several pathways, or RNA structures that can fold to low-energy structures that are close in energy to the ground state. By contrast the non-deterministic phenotypes that are more likely in a self-assembly context are of the kind described in [31], which often resemble uncontrolled protein aggregation. The number of potential different structures that can arise from a single genotype in such circumstances is extremely large. While limited non-determinism and plasticity play an important role in evolution, biological processes favour robust and therefore deterministic self-assembly in the majority of cases, such as the large variety of different protein complexes that have been observed in nature [32]. In the context of protein self-assembly, disease phenotypes are often the result of non-deterministic processes, such as protein aggregation. Sickle-cell disease [15] and the formation of amyloid fibrils [16] are prominent examples of this. It has been shown that GP map properties strongly determine evolutionary outcomes in RNA secondary structure [33]. The properties of non-deterministic self-assembly GP maps may therefore provide another perspective on the evolution of disease phenotypes.

ACKNOWLEDGMENTS

ST was supported by the EPSRC. SEA was supported by the Royal Society and the Gatsby Foundation.

[1] Peter Schuster, Walter Fontana, Peter F Stadler, and Ivo L Hofacker. From sequences to shapes and back: a case study in rna secondary structures. *Proceedings of the Royal Society of London B: Biological Sciences*, 255(1344):279–284, 1994.

[2] Massimo Pigliucci. Genotype–phenotype mapping and the end of the genes as blueprintmetaphor. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1540):557–566, 2010.

[3] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666, 1996.

[4] Erich Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73(5):2393, 1997.

[5] Iain G Johnston, Sebastian E Ahnert, Jonathan P K Doye, and Ard A Louis. Evolutionary dynamics in a simple model of self-assembly. *Physical Review E*, 83(6 Pt 2):066105–066105, May 2011.

[6] 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.

[7] A Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of The Royal Society B - Biological Sciences*, 275(1630):91–100, January 2008.

[8] Jacobo Aguirre, Javier M Buldú, Michael Stich, and Susanna C Manrubia. Topological structure of the space of phenotypes: the case of rna neutral networks. *PLOS One*, 6(10):e26324, 2011.

[9] 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.

[10] 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, March 2016.

[11] Andreas Wagner. Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity. *Biophysical Journal*, 106(4):955–965, February 2014.

[12] Olga Khersonsky Tawfik and Dan S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1):471–505, June 2010.

[13] Amir Aharoni, Leonid Gaidukov, Olga Khersonsky, Stephen McQ Gould, Cintia Roodveldt, and Dan S Tawfik. The 'evolvability' of promiscuous protein functions. *Nature Genetics*, 37(1):73, January 2005.

[14] Carlos Espinosa-Soto, Olivier C Martin, and Andreas Wagner. Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits. *Bmc Evolutionary Biology*, 11(1):5, January 2011.

[15] H Franklin Bunn. Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine*, 337(11):762–769, 1997.

[16] Samuel IA Cohen, Sara Linse, Leila M Luheshi, Erik Hellstrand, Duncan A White, Luke Rajah, Daniel E Otzen, Michele Vendruscolo, Christopher M Dobson, and Tuo-

mas PJ Knowles. Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism. *PNAS*, 110(24):9758–9763, 2013.

[17] S. E Ahnert, I. G Johnston, T. M. A Fink, J. P. K Doye, and A. A Louis. Self-assembly, modularity, and physical complexity. *Physical Review E*, 82(2):026117, August 2010.

[18] Andreas Wagner. *Robustness and evolvability in living systems*. Princeton University Press, 2013.

[19] Steffen Schaper. *On the significance of neutral spaces in adaptive evolution*. PhD thesis, University of Oxford, 2012.

[20] Pablo Catalán, Andreas Wagner, Susanna Manrubia, and José A Cuesta. Adding levels of complexity enhances robustness and evolvability in a multilevel genotype–phenotype map. *Journal of Royal Society Interface*, 15(138):20170516–10, January 2018.

[21] Giovanni Marco Dall'Olio, Jaume Bertranpetit, Andreas Wagner, and Hafid Laayouni. Human Genome Variation and the Concept of Genotype Networks. *PLoS ONE*, 9(6):e99424–11, June 2014.

[22] P Alberch. From genes to phenotype: dynamical systems and evolvability. *Genetica*, 84(1):5–11, 1991.

[23] Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1630):91–100, 2008.

[24] S F Greenbury and S E Ahnert. The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype–phenotype maps. *Journal of The Royal Society Interface*, 12(113):20150724, November 2015.

[25] Erich Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73(5):2393, 1997.

[26] Evandro Ferrada and Andreas Wagner. A comparison of genotype-phenotype maps for rna and proteins. *Biophysical journal*, 102(8):1916–1925, 2012.

[27] Marcel Weiß and Sebastian E Ahnert. Phenotypes can be robust and evolvable if mutations have non-local effects on sequence constraints. *Journal of Royal Society Interface*, 15(138):20170618–11, January 2018.

[28] L W Ancel and W FONTANA. Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 288(3):242–283, October 2000.

[29] Aditya Barve and Andreas Wagner. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461):203–206, July 2013.

[30] Nilesh Vaidya and Niles Lehman. One RNA plays three roles to provide catalytic activity to a group I intron lacking an endogenous internal guide sequence. *Nucleic Acids Research*, 37(12):3981–3989, July 2009.

[31] S. Tesoro and S. E. Ahnert. Nondeterministic self-assembly of two tile types on a lattice. *Phys. Rev. E*, 93:042412, Apr 2016.

[32] S. E. Ahnert, J. A. Marsh, H. Hernandez, C. V. Robinson, and S. A. Teichmann. A periodic table of protein complexes. *Science*, 350(6266):aaa2245–aaa2245, December 2015.

[33] 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, December 2015.