

## Research



**Cite this article:** García-Galindo P, Ahnert SE, Martin NS. 2023 The non-deterministic genotype–phenotype map of RNA secondary structure. *J. R. Soc. Interface* **20**: 20230132. <https://doi.org/10.1098/rsif.2023.0132>

Received: 8 March 2023

Accepted: 1 August 2023

### Subject Category:

Life Sciences–Physics interface

### Subject Areas:

biophysics, evolution, computational biology

### Keywords:

RNA, robustness, evolvability, many-to-many genotype–phenotype map

### Author for correspondence:

Paula García-Galindo

e-mail: [pg520@cam.ac.uk](mailto:pg520@cam.ac.uk)

<sup>†</sup>Present address: CRG (Barcelona

Collaboratorium for Modelling and Predictive Biology), Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.6778081>.

# The non-deterministic genotype–phenotype map of RNA secondary structure

Paula García-Galindo<sup>1</sup>, Sebastian E. Ahnert<sup>1,2</sup> and Nora S. Martin<sup>3,†</sup>

<sup>1</sup>Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK

<sup>2</sup>The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

<sup>3</sup>Rudolf Peierls Centre for Theoretical Physics, Beecroft Building, Parks Road, Oxford OX1 3PU, UK

PG-G, 0000-0003-0667-6098; SEA, 0000-0003-2613-0041; NSM, 0000-0003-4026-0985

Selection and variation are both key aspects in the evolutionary process. Previous research on the mapping between molecular sequence (genotype) and molecular fold (phenotype) has shown the presence of several structural properties in different biological contexts, implying that these might be universal in evolutionary spaces. The deterministic genotype–phenotype (GP) map that links short RNA sequences to minimum free energy secondary structures has been studied extensively because of its computational tractability and biologically realistic nature. However, this mapping ignores the phenotypic plasticity of RNA. We define a GP map that incorporates non-deterministic (ND) phenotypes, and take RNA as a case study; we use the Boltzmann probability distribution of folded structures and examine the structural properties of ND GP maps for RNA sequences of length 12 and coarse-grained RNA structures of length 30 (RNASHapes30). A framework is presented to study robustness, evolvability and neutral spaces in the ND map. This framework is validated by demonstrating close correspondence between the ND quantities and sample averages of their deterministic counterparts. When using the ND framework we observe the same structural properties as in the deterministic GP map, such as bias, negative correlation between genotypic robustness and evolvability, and positive correlation between phenotypic robustness and evolvability.

## 1. Introduction

Biological information is stored in our genome in DNA and RNA as linear sequences composed of a nucleotide quaternary alphabet [1]. In nature, RNA is commonly found as a single-stranded sequence that can fold, through nucleotide interactions, into a three-dimensional molecular structure. For short RNAs, this three-dimensional structure can be coarse-grained to a two-dimensional representation of the fold (secondary structure) [2]. The biological functions that RNA executes in the cell depend on its folded structure [2], and so a direct link between RNA sequence, its minimum free energy (MFE) fold, and its function can be constructed [3,4]. The sequence–structure mapping can be called the genotype–phenotype (GP) map for RNA, since the sequence is the biological information carrier (genotype) and the fold has a direct implication on function and therefore selection (phenotype) [4].

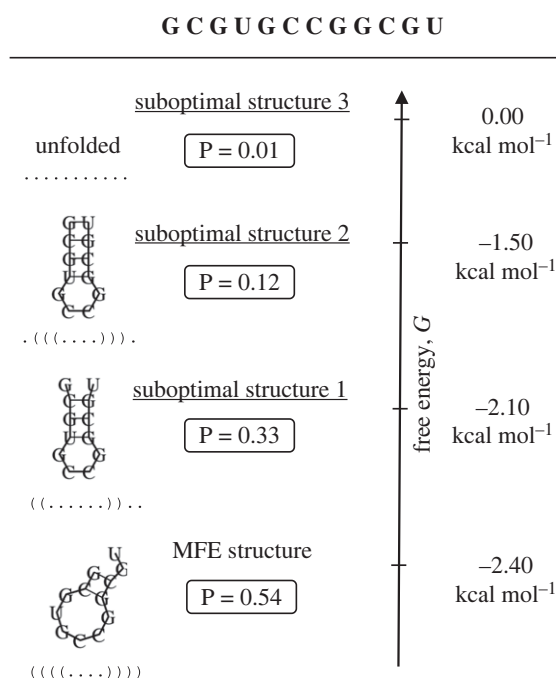
For RNA of length  $L$ , the GP map involves all possible  $L$ -sequences mapped to their respective MFE secondary structures, in which each sequence is connected to its point-mutation neighbours [5]. The GP map can then be interpreted as a road-map outlining the accessibility of phenotypes in terms of point mutations of genotypes [5]. Constructing this mapping is tractable for short RNAs with the currently available computational tools [6,7], which

is one of the main reasons why the GP map of RNA secondary structure has been studied so extensively [5].

The study of different GP maps which are not necessarily sequence-structure but represent a different phenotype (e.g. for proteins [8–10], metabolic networks [5] and gene regulatory networks (GRNs) [11]) has shown the prevalence of structural properties of the GP map that are common with the RNA sequence-structure map, suggesting that the GP maps exhibit properties that are characteristic of evolutionary spaces in general [1,5]. These properties include redundancy, bias, correlations between evolvability (i.e. the potential for phenotypic changes through mutations) and robustness (i.e. the probability of phenotype-preserving mutations), and the presence of neutral spaces and genetic correlations [5]. First, most GP maps display redundancy, which means that there are many genotypes mapping to a single phenotype, for example many sequences folding into a single structure in the RNA GP map. The set of genotypes mapping to the same phenotype is called a neutral space [1], or a neutral network if they are connected by point mutations [1,12]. Secondly, most GP maps are biased, which means that some phenotypes are much more redundant than others and thus have much larger neutral spaces. Thirdly, GP maps show a correlation between evolvability and robustness that is negative at genotypic level and positive at the phenotypic level [13]. Finally, GP maps usually display genetic correlations, which means that robustness is much higher than we would expect in a null model that simply accounts for neutral space sizes [14]. It is neutral spaces and networks in GP maps that explain the emergence of the more general concept of biological robustness [1,13].

In our discussions so far, and in most of the GP map literature more widely, GP maps are deterministic many-to-one mappings, in which each genotype maps to one phenotype. For RNA, the MFE fold represents this deterministic phenotype [1,5,15]. However, in nature, phenotypes can be non-deterministic (ND), either when the environment is variable, or even in a constant environment. For example, RNA folding is stochastic in a constant environment because of the inherent thermal fluctuations of the molecules at finite temperature [15]. Therefore, the nucleotide interactions create a range of possible folds described by the Boltzmann distribution. The structural phenotype of an RNA molecule is in fact ND because it is fluctuating in time through Boltzmann-distributed suboptimal structures [15]. Seen in another way, at any one instant, a large set of RNAs of the same genotype, placed in an environment of temperature  $T$ , will exhibit a range of folded structures, where the amount of genotypes with fold  $p$  will be proportional to the Boltzmann probability  $P(p|g) = \exp((G_{\text{ens}} - G_p)/kT)$ , where  $G_{\text{ens}}$  is the ensemble free energy ( $G_{\text{ens}} = -kT \cdot \ln Z$  with partition function  $Z$ ) and  $G_p$  is the free energy that corresponds to structure  $p$  [15] (cf. figure 1). Ance & Fontana's highly influential work referred to this as RNA plasticity, and studied its GP map [15]. They introduced an intrinsic feature of the plastic RNA GP map called plastogenetic congruence, which is a correlation between a genotype's suboptimal structures and the MFE structures of its point-mutation neighbours [15]. This feature was also shown to create a negative correlation between RNA plasticity and robustness [15].

Non-determinism is not only found in RNA, but is ubiquitous in biological systems [16]. Macromolecules such as proteins can also assume more than one folded structure



**Figure 1.** RNA plasticity example. Schematic of an example set of Boltzmann-distributed suboptimal structures for genotype GCGUGCCGGCGU with their corresponding free energies  $G$  and normalized probabilities  $P$ . The unfolded case is included as a suboptimal structure with characteristic free energy  $G = 0.0$  kcal mol<sup>-1</sup>. Secondary structures and energies are calculated using the ViennaRNA `subopt` command with default values [7].

[17]. Similarly, on the level of function, it has been shown that a single RNA sequence or a single enzyme can perform multiple functions [18–20]. Beyond the molecular scale, non-determinism also exists: genetically identical microbial cells can have varying phenotypes [21,22], for example due to stochastic gene expression [23]. Across scales, non-determinism could influence evolutionary processes. For example, studies have proposed that it can expedite evolution, but also lead to confinement to less evolvable sequences [15,24–27]. However, there are still many open questions about non-determinism's impact on evolution [28,29].

The strength of the GP map framework lies in its ability to clearly quantify the sequence-to-structure relationship [5]. The study of these maps has advanced greatly due to the rapid improvement of computational power and tools witnessed over recent decades [30]. But also particularly important, has been the formulation of a general quantifying framework. Quantities such as robustness and evolvability have been key to the exploration of how different deterministic GP maps compare and the finding of deterministic GP map's quasi-universal properties, which has eventually resulted in significant progress in understanding variation [5,30]. For the ND case, the work from Ance & Fontana [15] as well as further more recent work [31–36] that includes a ND phenotype in the context of GP maps, has improved our understanding but has not fully addressed how concepts such as robustness, evolvability and neutral spaces can be quantified for the ND case. We define a quantifying framework for non-determinism given by known phenotype probability distributions. In other words, we study the case of non-determinism as stochastic. We expect the ND GP map quantities to be consistent with averages of their deterministic counterparts, and that the framework can be a

baseline for further exploration of ND GP maps in general, of any biological context, which could help bring answers on the relationship between non-determinism and evolution.

The terms plasticity [15], promiscuity [30], intra-genotypic variability [37], non-determinism [5] and most recently, ‘probabilistic GP map (PrGP map)’ [35] have been used to describe many-to-many GP maps. Here we will use ‘ND’. The ND GP map we use is defined as a many-to-many mapping where the phenotypes of each genotype form a probability distribution, and for clarity we will refer to a general deterministic GP map, as ‘D GP map’. In the case of RNA, the MFE structure is usually taken to be the deterministic phenotype since it has the highest probability of folding and we will therefore compare our results from the full ND map with this MFE map. The work is structured as follows: first, we mathematically define ND quantities that are designed to correspond to averages of deterministic quantities over an ensemble. Then we validate this correspondence for both RNAs of  $L=12$  (conventional sequence-structure mapping) and  $L=30$  (RNASHAPES [38,39]). Following that, the deterministic quantities and their relationships are compared with the ND case to gain a deeper insight into the differences between the RNA ND GP map and the deterministic GP map. The computational methods used to construct the ND GP maps are explained at the end of this paper.

## 2. Results and discussion

In order to generalize GP map measurements such as robustness, evolvability and neutral space size to ND GP maps we take a stochastic approach. The stochastic versions of these measurements will quantify the averages of the corresponding deterministic quantities over the ND ensemble. This definition is general enough to be used in any biological context for which an ND GP map can be constructed. We use short RNA structures as a demonstration and validation of this approach.

### 2.1. Neutral spaces and robustness

In a deterministic GP map, the neutral set of a phenotype is the set of all genotypes that map to it. Its size as a fraction of genotype space is often referred to as the phenotypic frequency [13]. In the ND GP map, we can define neutral sets in terms of the average frequency  $\tilde{f}_p$  of a phenotype  $p$  over the ND probability distribution (as in [35,36,40]):

$$\tilde{f}_p = \frac{1}{K^L} \sum_{g \in \mathcal{G}_p} P(p|g), \quad (2.1)$$

where  $\mathcal{G}_p$  is the set of genotypes that contain the phenotype  $p$  in their ND ensemble and  $P(p|g)$  is the probability that genotype  $g$  gives phenotype  $p$ .  $K^L$  is the total number of genotypes in genotype space, where  $K$  is the genotype alphabet size and  $L$  is the genotype sequence length. For RNA, the ensemble is Boltzmann distributed and for sequences of length  $L$ , the total number of genotypes in genotype space is  $4^L$  since RNA has an alphabet of 4 nucleotides.

Robustness quantifies the insensitivity of a phenotype to genotypic mutations, and can be defined at the genotypic or phenotypic level [13]. In D GP maps, genotypic robustness  $\rho_g$  corresponds to the fraction of genotypic point-mutation

neighbours that leave the phenotype unchanged (also referred to as neutral neighbours) [13]

$$\rho_g = \frac{1}{(K-1)L} \sum_{g' \in \mathcal{N}_g} \delta_{g'p}, \quad (2.2)$$

where  $\mathcal{N}_g$  is the set of all point-mutation neighbours of  $g$ ,  $p$  is the phenotype of  $g$  and  $\delta_{g'p}$  equals 1 if the genotype  $g'$  maps to phenotype  $p$ , and 0 otherwise. The normalization  $(K-1)L$  is the total number of point-mutation neighbours in genotype space. In RNA, the number of such neighbours for any genotype is  $3L$ . If we want to generalize this definition to the ND case we need to consider the probability that  $g$  gives rise to phenotype  $p$  and the probability that genotypes in the mutational neighbourhoods of  $g$  also give rise to that phenotype. Thus, we define the ND genotypic robustness as

$$\tilde{\rho}_g = \frac{1}{(K-1)L} \sum_{p \in \mathcal{P}_g} P(p|g) \sum_{g' \in \mathcal{N}_g} P(p|g'), \quad (2.3)$$

where  $\mathcal{P}_g$  is the set of phenotypes  $p$  in the ND ensemble of genotype  $g$ , and  $\sum_{g' \in \mathcal{N}_g}$  is the sum over all point-mutation neighbours  $g'$  of  $g$ , which form the set  $\mathcal{N}_g$ . The maximum of  $\tilde{\rho}_g$  occurs when  $g$  only maps to one phenotype, and this phenotype is also the only phenotype for  $g'$ .

The deterministic phenotypic robustness  $\rho_p$  is the genotypic robustness (equation (2.2)) averaged over all genotypes that have phenotype  $p$  [1,13]

$$\rho_p = \frac{1}{K^L \cdot \tilde{f}_p} \sum_{g \in \mathcal{G}_p} \rho_g, \quad (2.4)$$

where  $\mathcal{G}_p$  is the neutral set of  $p$  with size  $|\mathcal{G}_p| = K^L \cdot \tilde{f}_p$ . For the ND phenotypic robustness, we formulate an average of the ND robustness of a specific phenotype  $p$  over the ND neutral set size, resulting in the following definition:

$$\tilde{\rho}_p = \frac{1}{K^L \cdot \tilde{f}_p \cdot (K-1)L} \sum_{g \in \mathcal{G}_p} P(p|g) \sum_{g' \in \mathcal{N}_g} P(p|g'), \quad (2.5)$$

where the ND neutral set size is expressed using equation (2.1), as  $K^L \cdot \tilde{f}_p$ . Equation (2.5) is analogous to the ND phenotype robustness definition recently proposed by Sappington & Mohanty [35].

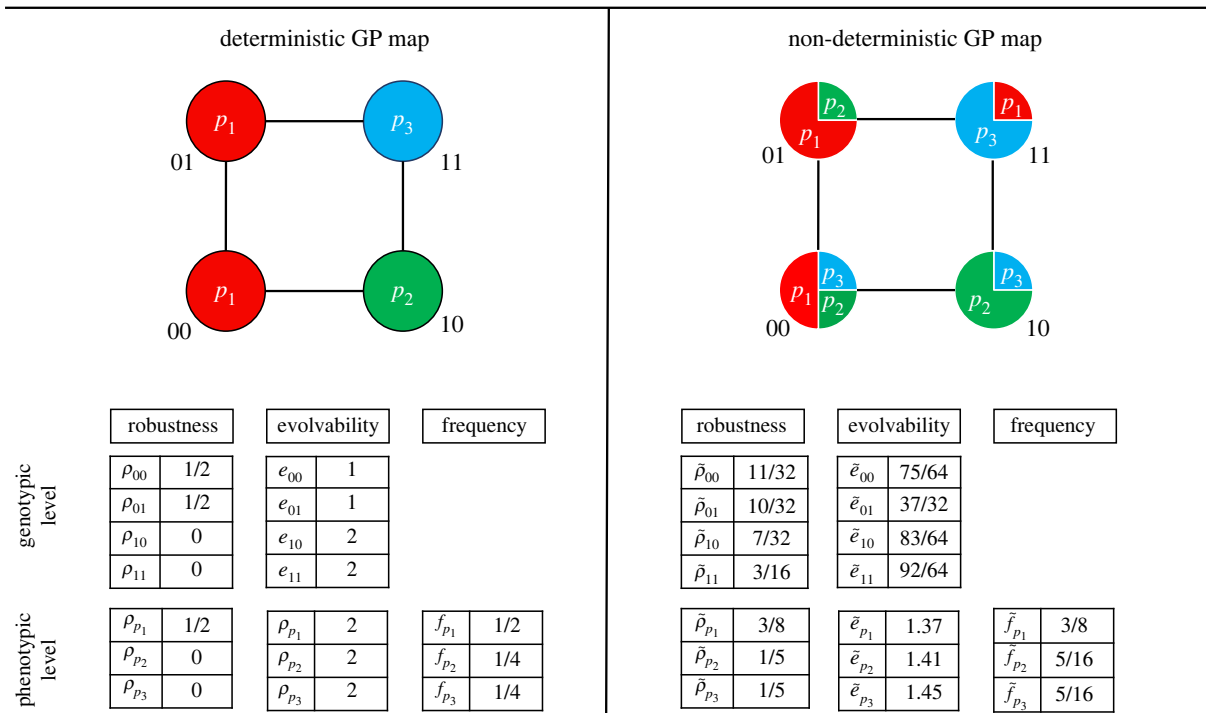
### 2.2. Evolvability

The capacity of a genotype or phenotype to produce phenotypic diversity is quantified by its ‘evolvability’ [13]. For the deterministic case, the genotypic evolvability of a given genotype  $g$  can be defined as the number of different structures found in its point-mutational neighbourhood [13]. The genotypic evolvability for the genotype  $g$  of phenotype  $p$  can therefore be written as

$$e_g = \sum_{p' \neq p} (1 - \prod_{g' \in \mathcal{N}_g} (1 - \delta_{g'p'})), \quad (2.6)$$

where, as before,  $\mathcal{N}_g$  is the set of all point-mutation neighbours of  $g$ , and  $\delta_{g'p'}$  equals 1 if the genotype  $g'$  maps to the phenotype  $p'$ , and 0 otherwise. For the ND case the  $\delta_{g'p'}$  is replaced by the probability  $P(p'|g')$ , resulting in an average of the number of different phenotypes in the neighbourhood of  $g$ . The average of each evolvability is weighted according to the probability that this phenotype is initially represented by  $g$

$$\tilde{e}_g = \sum_{p \in \mathcal{P}_g} P(p|g) \sum_{p' \neq p} (1 - \prod_{g' \in \mathcal{N}_g} (1 - P(p'|g'))), \quad (2.7)$$



**Figure 2.** Schematic depiction of simple deterministic and ND GP maps, and associated properties. For an extremely simple genotype space of size 4 (genotype length  $L = 2$  and alphabet size  $K = 2$ ) we illustrate the difference between deterministic (D) and non-deterministic (ND) GP maps. The relative area that a phenotype occupies in a genotype node represents the probability that this phenotype arises from this genotype in the ND phenotype ensemble. Each node in the deterministic mapping is, like for the RNA MFE GP map, the most probable phenotype in the ND case. The robustness and evolvability of genotypes and phenotypes, as well as the phenotypic frequency are calculated for both the deterministic and ND case using equations (2.1)–(2.9). A detailed step-by-step calculation of the quantities is shown in the electronic supplementary material, S1.

where the sum  $\sum_{p' \neq p}$  is over all possible phenotypes  $p'$  in the ND GP map except  $p$ .

The phenotypic evolvability of a phenotype  $p$  counts all different phenotypes in the mutational neighbourhood of  $p$  as a whole, or in other words, of all genotypes with phenotype  $p$  [13]. For the deterministic case, we can write this as

$$e_p = \sum_{p' \neq p} (1 - \prod_{g \in G_p} \prod_{g' \in \mathcal{N}_g} (1 - \delta_{g'p'})). \quad (2.8)$$

Note the additional product compared with equation (2.6) to make sure we are considering all  $g$ s with  $p$ , represented by the neutral set  $G_p$ . The ND phenotypic evolvability  $\tilde{e}_p$  represents the average number of accessible phenotypes from  $p$ , meaning the set of genotypes in  $\mathcal{G}_p$ . It can be formulated by replacing the  $\delta_{g'p'}$  in equation (2.8) with the product of probabilities  $P(p'|g')P(p|g)$  and accounting for all genotypes in  $\mathcal{G}_p$

$$\tilde{e}_p = \sum_{p' \neq p} (1 - \prod_{g \in \mathcal{G}_p} \prod_{g' \in \mathcal{N}_g} (1 - P(p'|g')P(p|g))). \quad (2.9)$$

A simple GP map schematic such as in figure 2 exemplifies the difference between the deterministic and ND case. The schematic shows the deterministic GP map as the map that assigns the most probable phenotype from the ND map for each genotype, which is analogous to how the MFE GP map of RNA relates to its ND GP map.

### 2.3. Validation of non-deterministic quantities

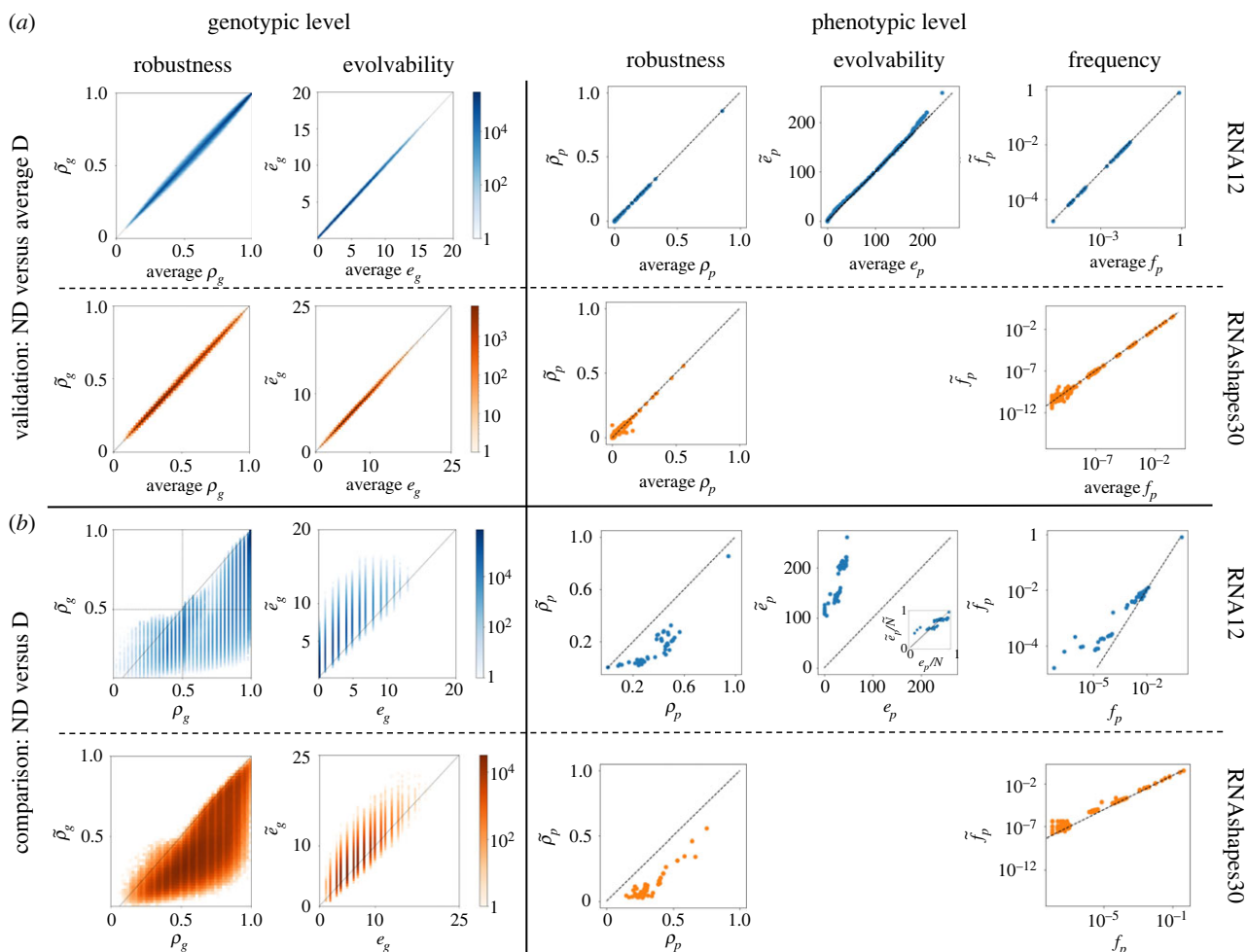
We have defined the ND quantities in equations (2.1), (2.3), (2.5), (2.7) and (2.9) to give us the averages of the deterministic quantities over a large number of realizations of the phenotype

ensemble. To validate these ND definitions computationally we use short RNA, a biologically realistic computational GP map for molecular evolution. The ND GP maps for RNA of  $L = 12$  and RNASHAPES of  $L = 30$  [39] are computationally constructed using the ViennaRNA package (see Methods). The validation consists of comparisons between the ND quantities for the ND GP map and the deterministic quantities averaged over a large sample ( $N = 500$ ) of GP map realizations: each realization is produced by drawing a sample from the Boltzmann distribution of each genotype, producing a single phenotype for each genotype in a given realization. As can be seen in figure 3a the ND definitions of robustness, evolvability and frequency match the averages of the deterministic quantities as expected, both at the genotypic and phenotypic levels. The phenotypic evolvability cannot be calculated with sampling methods as discussed in [41], therefore  $e_p$  and  $\tilde{e}_p$  are not computed for RNASHAPES30.

### 2.4. Comparison of non-deterministic genotype–phenotype map properties with their minimum free energy counterparts

We next compare the ND quantities in the ND GP map with their deterministic MFE counterparts. The latter are the GP map quantities that have been studied extensively in the RNA GP map literature (see Introduction). Figure 3b shows these comparisons.

At the genotypic level, figure 3b shows similar behaviour for both RNA12 and RNASHAPES30. We find that genotypes with high genotypic robustness in the MFE GP map tend to



**Figure 3.** Non-deterministic framework validation and comparison. (a) The definitions of non-deterministic robustness  $\tilde{\rho}$ , evolvability  $\tilde{e}$  and phenotypic frequency  $\tilde{f}_p$  quantities are validated by comparing them with the averages of the equivalent deterministic quantities for RNA12 and RNASHapes30, across 500 samples of the ND GP map in both cases. (b) Direct comparison between non-deterministic quantities on the ND GP map and the deterministic quantities on the MFE GP map. The phenotypic evolvability cannot be calculated through sampling methods [41], therefore we do not present evolvability plots for RNASHapes30. The phenotypic evolvability comparison for RNA12 includes an inset for normalized phenotypic evolvability  $e_p/N$ ,  $\tilde{e}_p/\tilde{N}$ , which shows a very close correspondence of these two quantities (see full-sized figure in electronic supplementary material).

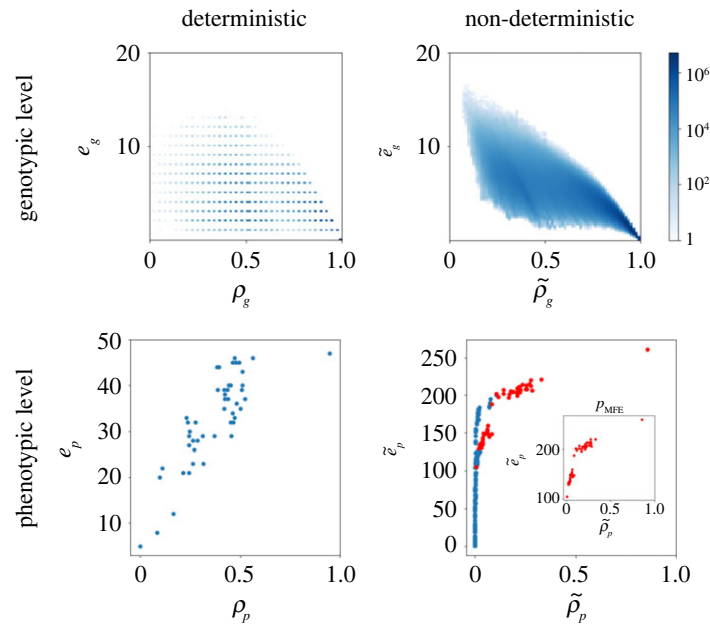
also have high robustness in the ND GP map. However, we find that in most cases the D genotypic robustness is higher than the ND genotypic robustness ( $\rho_g \geq \tilde{\rho}_g$ ), except for a low-density section of genotypes with  $\tilde{\rho}_g > \rho_g$  in the regime  $\tilde{\rho}_g, \rho_g < 0.5$ . Similarly, genotypes with high genotypic evolvability in the MFE GP map tend to also have high evolvability in the ND GP map. However, genotypic evolvability is higher in the ND case (i.e.  $\tilde{e}_g \geq e_g$ ) for the majority of genotypes. Therefore, we conclude that (most) genotypes in the ND GP maps of RNA secondary structure will on average be less robust and more evolvable than the corresponding MFE GP map.

At the phenotypic level, figure 3b shows each corresponding quantity for the same phenotype. In the RNA12 ND GP map (energy range  $15k_B T$ , see Methods) there exist a total of  $\tilde{N} = 271$  phenotypes, while in the deterministic case there are only  $N = 48$ . Therefore, we can only compare the GP map quantities for the 48 phenotypes present in both maps. Given the result at the genotypic level, it is not surprising that phenotypic robustness is lower in the ND case than in the deterministic one. Phenotypic evolvability, however, shows a sharp increase, meaning that the same phenotypes are surrounded by many more phenotypes. However, this is a function of the increased total number of phenotypes

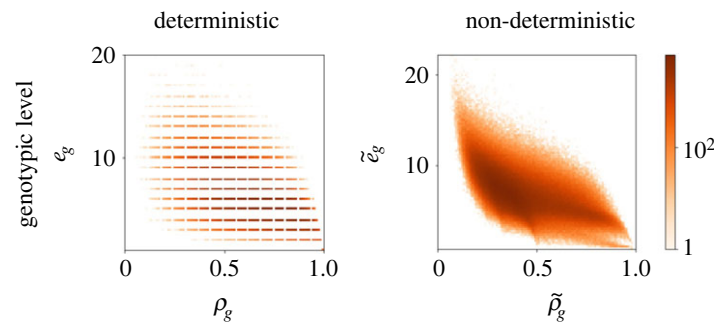
found in the ND GP map, as shown by an inset plot comparing the fractions of all possible phenotypes that are discovered in the neighbourhood of a given phenotype  $p$ , which we will refer to as the normalized phenotypic evolvabilities  $\tilde{e}_p/\tilde{N}$  (for the ND GP map) and  $e_p/N$  (for the D GP map). These normalized evolvabilities are similar in the MFE and ND GP maps (inset), but phenotypes at the low-evolvability end have slightly higher normalized evolvabilities in the ND case and thus a higher fraction of phenotypes is accessible from them. In all phenotypic plots for RNA12 the outlier at the top right of the distribution represents the unfolded structure, which has a disproportionately large neutral space [14]. The phenotypic frequencies exhibit  $\tilde{f}_p > f_p$  for the most part, particularly for low values of  $f_p$  and  $\tilde{f}_p$ . This means that low-frequency phenotypes have larger neutral spaces in the ND GP map than in the MFE GP map, for both RNA12 and RNASHapes30.

## 2.5. Structural properties of the non-deterministic genotype–phenotype map

Next we determine whether the structural properties that have been observed across several GP maps, including the RNA MFE GP map (see Introduction) also hold in the ND GP map. First, the so-called ‘bias’ observed in many GP



**Figure 4.** RNA12 robustness versus evolvability. At the genotypic level evolvability and robustness exhibit negative correlations, with a Pearson correlation coefficient of  $-0.73$  in the MFE GP map, and  $-0.85$  in the ND GP map. At the phenotypic level, the correlation between evolvability and robustness is positive with Pearson correlation coefficients  $0.81$  (MFE GP map) and  $0.61$  (ND GP map). The structures that appear in both the MFE and ND GP map are highlighted in red in the ND phenotypic evolvability/robustness plot.



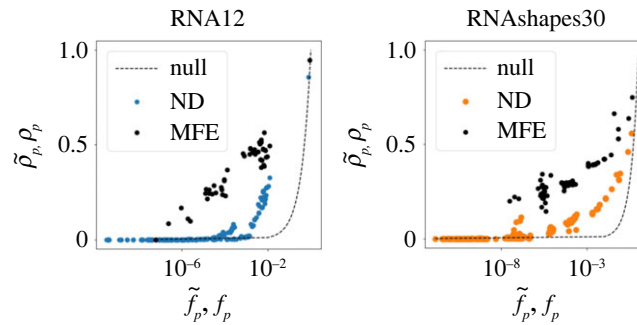
**Figure 5.** RNASHAPES30 robustness versus evolvability. The MFE GP map shows a negative correlation between the genotypic robustness  $\rho_g$  and genotypic evolvability  $e_g$ , with a Pearson coefficient of  $-0.56$ . A negative correlation is also observed in the ND GP map with Pearson coefficient of  $-0.64$ .

maps describes the highly skewed distribution of phenotypic frequencies [5]. In most GP maps, the vast majority of genotypes maps to a small number of phenotypes, and the remainder maps to many different phenotypes. This property is also found in the ND RNA GP maps because the phenotypic frequency values,  $\tilde{f}_p$ , differ by orders of magnitude, as shown in figure 3 (as found in [35,36,40]). Second, many GP maps, including the RNA MFE GP map, exhibit a negative correlation of genotypic evolvability and robustness and a positive correlation of phenotypic evolvability and robustness [13]. The negative correlation in the genotypic case is expected as the neighbourhood of an individual genotype cannot contain many instances of the same phenotype and also instances of many other phenotypes. A trade-off is therefore inevitable. The positive correlation at the phenotypic level, however, is a consequence of the shape and size of neutral networks, as well as the organization of information into biological sequences [13,42,43]. These relationships also hold in the case of the ND GP map, as can be seen in figure 4 for RNA12. At the genotypic level, the expected MFE GP map negative correlation is present (Pearson correlation coefficient of  $-0.73$ ) and a similar negative correlation can be observed in the ND GP map (Pearson

correlation coefficient of  $-0.85$ ). At the phenotypic level, the expected positive correlation is observed for the MFE GP map (Pearson correlation coefficient of  $0.81$ ). The ND GP map also shows a positive correlation (Pearson correlation coefficient of  $0.61$ ), but the points are somewhat differently distributed. If we distinguish MFE phenotypes (highlighted in red, and also plotted separately in the inset) among all phenotypes in the ND GP map, we recover a distribution that mirrors the MFE case more closely. The phenotypes that only appear in the ND GP map form a long tail with very low robustnesses but a wide range of evolvability values.

In figure 5, the RNASHAPES30 display a similar relationship between evolvability and robustness at the genotypic level, in the form of negative correlations for both the MFE and ND GP maps (Pearson coefficient of  $-0.56$  for the MFE GP map and  $-0.64$  for the ND GP map). Again, because the phenotypic evolvability cannot be calculated with sampling methods [41], we do not study the correlations at the phenotypic level for RNASHAPES30. These results indicate that the commonly observed correlations between evolvability and robustness are also present in the ND GP map.

Many GP maps, including the RNA MFE GP map, display neutral correlations [14]. One might expect these to



**Figure 6.** The presence of neutral correlations: phenotypic robustness versus phenotypic frequency for RNA12 and RNAs30. Neutral correlations are still present in the ND GP maps for high phenotype frequencies,  $\tilde{\rho}_p \propto \log \tilde{f}_p$ . The robustnesses of low-frequency phenotypes equal those of the null model,  $\tilde{\rho}_p = \tilde{f}_p$ . In the MFE GP maps, the neutral correlations are present for all frequencies,  $\rho_p \propto \log f_p$ , as previously observed [14].

also be found in the ND GP map, due to plastogenetic congruence. This concept, introduced by Ancel & Fontana [15] relates mutational neighbourhood of a structure to its Boltzmann distribution (see electronic supplementary material). To verify that the ND GP map also contains neutral correlations we verify whether  $\tilde{\rho}_p \propto \log \tilde{f}_p$  holds, as has been found for D GP maps [4,10,14]. Figure 6 shows the relationship between these two quantities, and verifies the presence of correlations, as for most phenotypes the robustness  $\tilde{\rho}_p$  is significantly larger than the null model representing a random distribution of phenotypes with  $\tilde{\rho}_p = \tilde{f}_p$ , as previously found for the D and recently also the ND GP maps [14,35]. We also observe, in both the RNA12 and RNAs30 ND GP maps, the presence of the recently observed biphasic robustness scaling [35]: for higher-frequency phenotypes, neutral correlations are present  $\tilde{\rho}_p \propto \log \tilde{f}_p$  but suppressed compared with the MFE GP map, which follows from the lower phenotypic robustness found in figure 3; in lower-frequency phenotypes, the correlations disappear and the relationship becomes analogue to the null model  $\tilde{\rho}_p \propto \tilde{f}_p$  [35].

### 3. Conclusion

The study of GP maps and their structural properties provides an important perspective on evolutionary processes that complements the widely studied concept of the fitness landscape [30,44]. For example, a determinant feature that has driven the evolution of natural functional RNAs is the accessibility of the phenotypes in the GP map which depends on the neutral networks and their distribution [5]. This suggests that the GP map structure and its implications for variation can be just as important as selection pressures [5], as new phenotypes need to be accessible in order to be selected. Robustness, evolvability and neutral spaces have become definitions in the general framework to quantify the relationship between the genotype and phenotype in GP maps, and have been key to the exploration of the quasi-universal structural properties seen in GP maps in different biological contexts [1,45]. Our contribution is a generalization of this framework to the ND GP map, for which each genotype has an ensemble of possible phenotypes that form a probability distribution. In molecular evolution, non-determinism plays a fundamental role since molecules are exposed to thermal fluctuations and produce ND phenotypes in the form of a statistical distribution of folded structures. The study of ND GP maps therefore provides a more realistic picture of variation and its impact on

molecular evolution. Such a ND framework can also be applied on the functional level and beyond the molecular scale. For example, there exist experimental observations of enzyme promiscuity [20], or of microbial ND phenotypes [21–23]. The results discussed in this paper provide a foundation for further research on ND GP maps and future work should apply the quantities defined here to the full ND version of other widely studied GP maps, for example proteins or GRNs. A particularly interesting avenue to explore would be how these ND quantities impact evolutionary dynamics on the ND GP maps. In the simplest models, the ND fitness can be computed as the fitness average over the ND phenotype ensemble [28]. The impact of this many-to-many fitness landscape on evolutionary trajectories and how functional structures have evolved is an open question of biological interest. In addition, further work should build on Ancel and Fontana's work [15] and address the error-threshold problem for a range of structures on the full RNA ND GP map. The ND characteristics we have defined here provide a foundation for such a research agenda.

## 4. Methods

### 4.1. RNA12

The ND GP map is constructed by computing the Boltzmann ensemble of suboptimal structures and their probabilities for each RNA sequence of length  $L$ . For the case of RNA  $L = 12$ , the folds are calculated for every sequence through ViennaRNA with all parameters set to default values (e.g. the temperature  $T = 37^\circ\text{C}$ ), using the ViennaRNA suboptimal function [7] and the Boltzmann probabilities of these are obtained using the partition function. The energy range for the suboptimals is  $15k_B T$  as in [46] to be consistent with the RNAs30 data. The final ND GP map is constructed by mapping each genotype sequence to its ensemble of structures in the energy range (including unfolded structure), as well as their respective normalized probabilities.

### 4.2. RNAs30 for length $L = 30$

The RNAs30 calculations rely on the implementation in [46], which uses the ViennaRNA package (v. 2.4.14) [7]. This is faster than the original RNAs30 program for short sequences and thus allows us to work with larger samples of sequences and structures. We use the same parameters as in [46] (shape level 2, temperature  $T = 37^\circ\text{C}$ , energy range  $15 k_B T$ , sequence length  $L = 30$ , no isolated base pairs). Unlike [46], however, we compute the deterministic map by taking the most frequent structure in the Boltzmann ensemble as the deterministic folded structure, without additional requirements on its relative frequency.

Since there are  $4^{30} \approx 10^{18}$  genotypes of sequence length  $L = 30$ , we have to rely on sampling: we simply work with  $10^6$  randomly generated sequences ( $10^5$  for the average case, which is computationally more expensive due to the necessary repetitions). For the plots showing genotype robustness and evolvability, we simply plot one value for each sequence in the sample—thus, not all possible genotypes are in the sample, but we obtain an overview of common genotypic robustness and evolvability values for a large random sample of genotypes. For the phenotypic robustness estimates, we simply approximate each sum over all sequences by a sum over our sequence sample. This approach was tested for RNA12, where we have exact data as a reference (shown in the electronic supplementary material, section S1.1). To investigate whether the chosen sequence sample is large enough, we analysed how robust our results are to sub-sampling (shown in the electronic supplementary material, section S1.2). Phenotypic evolvabilities are not estimated since they cannot be inferred reliably from samples with current techniques [41].

Phenotypic frequencies were also estimated from a random sequence sample. However, phenotypic frequency calculations are faster than robustness calculations because they do not require us to examine the mutational neighbours of each

sequence, and so we used a larger sample of  $10^8$  sequences. Note that our reliance on sampling methods means that quantities for low-frequency phenotypes, which do not appear in the sample, cannot be estimated.

**Ethics.** This work did not require ethical approval from a human subject or animal welfare committee.

**Data accessibility.** Access code and data from the GitHub repository: <https://github.com/paulagrcg/The-non-deterministic-genotype-phenotype-map-of-RNA-secondary-structure>.

**Declaration of AI use.** We have not used AI-assisted technologies in creating this article.

**Authors' contributions.** P.G.-G.: conceptualization, data curation, formal analysis, visualization, writing—original draft, writing—review and editing; S.E.A.: conceptualization, writing—review and editing; N.S.M.: conceptualization, data curation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

**Conflict of interest declaration.** We declare we have no competing interests.

**Funding.** P.G.-G. was funded by a “La Caixa” Foundation postgraduate scholarship and N.S.M. acknowledges funding from the Issachar Fund.

## References

- Wagner A. 2013 *Robustness and evolvability in living systems*, vol. 15. Princeton, NJ: Princeton University Press.
- Cowperthwaite MC, Meyers LA. 2007 How mutational networks shape evolution: lessons from RNA models. *Annu. Rev. Ecol. Evol. Syst.* **38**, 203–230. (doi:10.1146/annurev.ecolsys.38.091206.095507)
- Fontana W, Schuster P. 1998 Shaping space: the possible and the attainable in RNA genotype–phenotype mapping. *J. Theor. Biol.* **194**, 491–515. (doi:10.1006/jtbi.1998.0771)
- Aguirre J, Buldú JM, Stich M, Manrubia SC. 2011 Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS ONE* **6**, e26324. (doi:10.1371/journal.pone.0026324)
- Ahnert SE. 2017 Structural properties of genotype–phenotype maps. *J. R. Soc. Interface* **14**, 20170275. (doi:10.1098/rsif.2017.0275)
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994 Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188. (doi:10.1007/BF00818163)
- Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011 ViennaRNA Package 2.0. *Algor. Mol. Biol.* **6**, 26. (doi:10.1186/1748-7188-6-26)
- Bornberg-Bauer E. 1997 How are model protein structures distributed in sequence space?. *Biophys. J.* **73**, 2393–2403. (doi:10.1016/S0006-3495(97)78268-7)
- Ahnert SE, Johnston IG, Fink TMA, Doye JPK, Louis AA. 2010 Self-assembly, modularity, and physical complexity. *Phys. Rev. E* **82**, 026117. (doi:10.1103/PhysRevE.82.026117)
- Greenbury SF, Johnston IG, Louis AA, Ahnert SE. 2014 A tractable genotype–phenotype map modelling the self-assembly of protein quaternary structure. *J. R. Soc. Interface* **11**, 20140249. (doi:10.1098/rsif.2014.0249)
- Catalán P, Manrubia S, Cuesta JA. 2020 Populations of genetic circuits are unable to find the fittest solution in a multilevel genotype–phenotype map. *J. R. Soc. Interface* **17**, 20190843. (doi:10.1098/rsif.2019.0843)
- Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994 From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **255**, 279–284. (doi:10.1098/rspb.1994.0040)
- Wagner A. 2008 Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B* **275**, 91–100. (doi:10.1098/rspb.2007.1137)
- Greenbury SF, Schaper S, Ahnert SE, Louis AA. 2016 Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLoS Comput. Biol.* **12**, e1004773. (doi:10.1371/journal.pcbi.1004773)
- Ancel LW, Fontana W. 2000 Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288**, 242–83. (doi:10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-0)
- Tawfik DS. 2010 Messy biology and the origins of evolutionary innovations. *Nat. Chem. Biol.* **6**, 692–696. (doi:10.1038/nchembio.441)
- Porter LL, Looger LL. 2018 Extant fold-switching proteins are widespread. *Proc. Natl Acad. Sci. USA* **115**, 5968–5973. (doi:10.1073/pnas.1800168115)
- Schultes EA, Bartel DP. 2000 One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448–452. (doi:10.1126/science.289.5478.448)
- Vaidya N, Lehman N. 2009 One RNA plays three roles to provide catalytic activity to a group I intron lacking an endogenous internal guide sequence. *Nucleic Acids Res.* **37**, 3981–3989. (doi:10.1093/nar/gkp271)
- Khersonsky O, Tawfik DS. 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505. (doi:10.1146/annurev-biochem-030409-143718)
- Veening J-W, Smits WK, Kuipers OP. 2008 Bistability, epigenetics, and bet-hedging in bacteria. *Annu. Rev. Microbiol.* **62**, 193–210. (doi:10.1146/annurev.micro.62.081307.163002)
- Ackermann M. 2015 A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol.* **13**, 497–508. (doi:10.1038/nrmicro3491)
- Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002 Stochastic gene expression in a single cell. *Science* **297**, 1183–1186. (doi:10.1126/science.1070919)
- Baldwin JM. 1896 A new factor in evolution. *Am. Nat.* **30**, 441–451. (doi:10.1086/276408)
- Ancel LW. 2000 Undermining the Baldwin expediting effect: does phenotypic plasticity accelerate evolution? *Theor. Popul. Biol.* **58**, 307–319. (doi:10.1006/tpbi.2000.1484)
- Hendry AP. 2015 Key questions on the role of phenotypic plasticity in eco-evolutionary dynamics. *J. Hered.* **107**, 25–41. (doi:10.1093/jhered/esv060)
- Espinosa-Soto C, Martin OC, Wagner A. 2011 Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits. *BMC Evol. Biol.* **11**, 5. (doi:10.1186/1471-2148-11-5)
- Aguirre J, Catalán P, Cuesta JA, Manrubia S. 2018 On the networked architecture of genotype spaces and its critical effects on molecular evolution. *Open Biol.* **8**, 180069. (doi:10.1098/rsob.180069)
- Aguirre J. 2022 Life finds a way. *Nat. Ecol. Evol.* **6**, 1599–1600. (doi:10.1038/s41559-022-01877-x)
- Manrubia S *et al.* 2021 From genotypes to organisms: state-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Phys. Life Rev.* **38**, 55–106. (doi:10.1016/j.plev.2021.03.004)



31. Wagner A. 2014 Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity. *Biophys. J.* **106**, 955–965. (doi:10.1016/j.bpj.2014.01.003)
32. Rezaadegan R, Reidys C. 2018 Degeneracy and genetic assimilation in RNA evolution. *BMC Bioinf.* **19**, 543. (doi:10.1186/s12859-018-2497-3)
33. Jouffrey V, Leonard AS, Ahnert SE. 2021 Gene duplication and subsequent diversification strongly affect phenotypic evolvability and robustness. *R. Soc. Open Sci.* **8**, 201636. (doi:10.1098/rsos.201636)
34. Martin NS, Ahnert SE. 2022 Fast free-energy-based neutral set size estimates for the RNA genotype–phenotype map. *J. R. Soc. Interface* **19**, 20220072. (doi:10.1098/rsif.2022.0072)
35. Sappington A, Mohanty V. 2023 Probabilistic genotype-phenotype maps reveal mutational robustness of RNA folding, spin glasses, and quantum circuits. *arXiv*. (doi:10.48550/arXiv.2301.01847)
36. Martin NS, Ahnert SE. 2023 The Boltzmann distributions of folded molecular structures predict likely changes through random mutations. *bioRxiv*. (doi:10.1101/2023.02.22.529545)
37. Bruijning M, JE. Metcalf C, Jongejans E, Ayroles JF. 2020 The evolution of variance control. *Trends Ecol. Evol.* **35**, 22–33. (doi:10.1016/j.tree.2019.08.005)
38. Giegerich R, Voß B, Rehmsmeier M. 2004 Abstract shapes of RNA. *Nucleic Acids Res.* **32**, 4843–4851. (doi:10.1093/nar/gkh779)
39. Voß B, Giegerich R, Rehmsmeier M. 2006 Complete probabilistic analysis of RNA shapes. *BMC Biol.* **4**, 5. (doi:10.1186/1741-7007-4-5)
40. Dingle K, Ghaddar F, Šulc P, Louis AA. 2021 Phenotype bias determines how natural RNA structures occupy the morphospace of all possible shapes. *Mol. Biol. Evol.* **39**, msab280. (doi:10.1093/molbev/msab280)
41. Weiß M. 2020 Structural properties of genotype-phenotype maps: models, estimates and evolution. PhD thesis, University of Cambridge.
42. Greenbury SF, Ahnert SE. 2015 The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype–phenotype maps. *J. R. Soc. Interface* **12**, 20150724. (doi:10.1098/rsif.2015.0724)
43. Weiß M, Ahnert SE. 2018 Phenotypes can be robust and evolvable if mutations have non-local effects on sequence constraints. *J. R. Soc. Interface* **15**, 20170618. (doi:10.1098/rsif.2017.0618)
44. Greenbury SF, Louis AA, Ahnert SE. 2022 The structure of genotype-phenotype maps makes fitness landscapes navigable. *Nat. Ecol. Evol.* **6**, 1742–1752. (doi:10.1038/s41559-022-01867-z)
45. Wagner A, Fell DA. 2001 The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**, 1803–1810. (doi:10.1098/rspb.2001.1711)
46. Martin NS, Ahnert SE. 2021 Insertions and deletions in the RNA sequence-structure map. *J. R. Soc. Interface* **18**, 20210380. (doi:10.1098/rsif.2021.0380)