Double-replica theory for evolution of genotype-phenotype interrelationship

Tuan Minh Pham [1] and Kunihiko Kaneko [1,2]

[1]*The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen 2100-DK, Denmark*

[2]*Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo, Komaba, Tokyo 153-8902, Japan*

(Received 28 November 2022; accepted 22 March 2023; published 21 April 2023)

The relationship between genotype and phenotype plays a crucial role in determining the function and robustness of biological systems. Here the evolution progresses through the change in genotype, whereas the selection is based on the phenotype, and the genotype-phenotype relation also evolves. The theory for such phenotypic evolution remains poorly developed, in contrast to evolution under the fitness landscape determined by genotypes. Here we provide a statistical-physics formulation of this problem by introducing replicas for genotype and phenotype. We apply it to an evolution model in which phenotypes are given by spin configurations; genotypes are an interaction matrix for spins to give the Hamiltonian, and the fitness depends only on the configuration of a subset of spins called the target. We describe the interplay between the genetic variations and phenotypic variances by noise in this model by our approach that extends the replica theory for spin glasses to include a spin replica for phenotypes and a coupling replica for genotypes. Within this framework we obtain a phase diagram of the evolved phenotypes against the noise and selection pressure, where each phase is distinguished by the fitness and overlaps for genotypes and phenotypes. Among the phases, a robust fitted phase, relevant to biological evolution, is achieved under the intermediate level of noise (temperature), where robustnesses to noise and to genetic mutation are correlated, as a result of replica symmetry. We also find a tradeoff between maintaining a high fitness level of phenotype and acquiring a robust pattern of genes as well as the dependence of this tradeoff on the ratio between the size of the functional (target) part to that of the remaining nonfunctional (non-target) one. The selection pressure needed to achieve high fitness increases with the fraction of target spins.

DOI: 10.1103/PhysRevResearch.5.023049

I. INTRODUCTION

Evolution under given fitness landscape in the space of genotypes has been studied extensively [1,2]. Here genotypes are changed (mutated) in the reproduction process, and those that have higher fitness, i.e., higher offspring-production rate, are selected to the next generation. The fitness, however, is not directly determined by the genes, but rather by phenotypes, that are the characteristics of a biological system and are also modulated by environmental effects, such as thermal noise [3]. Examples of phenotypes include the shape of folded proteins, a set of intracellular chemical concentrations, specific functions of an organism, and so forth. These phenotypes are shaped as a result of dynamical process, whose rule is determined by genes. The evolution of phenotypes is thus shaped by the genetic evolution.

In addition to the variations of phenotypes induced by genetic mutations during the evolution, phenotypes of isogenic individuals are also generally variable under noise, resulting in their stochastic dynamics. Cells involve stochastic gene expression dynamics, whereas protein folding dynamics to give the protein shape is under thermal noise [4–7]. Fitted phenotypic states hence are better preserved under noise, i.e., they keep robustness to noise, while robustness to mutation will also be required. The achievement of robustness of phenotypes to noise and to mutation is important to evolution, as was discussed recently [8–11]. Now, considering stochastic dynamics of phenotypes, a general formulation of the evolution of such genotype-phenotype mapping and phenotypic robustness is hence wanted.

Underlying such stochastic dynamics are the interactions among a vast number of elements that constitute a biological system. A cell consists of a huge variety of interacting molecules, and its constituent polymers (proteins) are composed of many monomers (residues). Statistical physics [12,13] provides an appropriate description of the states of such interacting elements under noise, and hence it can be adopted to yield a proper formulation of the mapping from genotype to phenotype as a dynamical process that shapes the phenotypes. To this kind of study, use of spin models is relevant, as it has been extensively analyzed [14–16]. Here phenotypes are regarded as spin configurations that are updated by a Hamiltonian with spin-spin interactions under thermal noise, whereas genotypes specify such spin-spin interactions. Individuals are subjected to selection pressure, where fitness, the number of selected offspring, is given by a function of configuration of a subset of spins, termed target spins. It is then important to identify possible phases

for genotype-phenotype mapping by using the set of order parameters, a concept rooted in statistical physics.

In the present paper, to address these problems in a systematic way, we formulate double replica methods both for spins (phenotypes) and couplings (genotypes), as well as both for target and nontarget parts. Even though we adopted spin-coupling representation here our formulation can generally be applied to other problems, in which the interactions among many degrees of freedom are also dynamical variables with their own dynamics. For the sake of demonstration, however, we here explain this approach by using specifically a spin-glass Hamiltonian model developed in [14–16]. These works uncovered that, within an intermediate range of phenotypic noise and under sufficiently strong selection pressure, the evolved spin-systems could attain high fitness and robustness to noise and mutation. A systematic way to elucidate the condition to achieve such robust fitted states with regards to the noise strength, selection pressure, and relative size of target spins, and to understand possible relationship between robustness to mutation and to noise, however, has not been developed yet.

One might expect an application of mean-field methods for disordered systems [17] in this model since a spin-glass Hamiltonian formulation was adopted by replacing random couplings among spins (phenotypes) by genotypes that are evolving [14–16]. Gradual change in the couplings might fit with partial annealing approach based on a finite number of replicas n. However, such study is restricted to the case in which the coupling dynamics are affected by a spin-spin correlation term [18–22], and is not directly applicable for our purpose, in which the coupling dynamics depend on the fitness determined by the spin configurations. Another theoretical method assumes the quench limit ($n \to 0$) for a replicated spin system [23], in which the couplings are treated effectively as "static" and hence is not suitable to investigate the evolution of both genotypes (couplings) and phenotypes (spin configurations).

In this paper we develop a mean-field approach that we term double replica theory. It describes the evolution of both genotypes and phenotypes by considering spins and links as two different replica species. With this formulation, we obtain fitness and replica overlaps for spins and couplings, which work as the order parameters. Using these order parameters we identify five regions in the temperature vs selection pressure phase diagram: two nonfitted paramagnetic phases, fitted and nonfitted spin glass phases, and a robust fitted phase. The last phase is the most biologically important, and can only be achieved under intermediate noise level (temperature) and sufficient selection pressure, whereas robustness can only arrive at the cost of lowering the fitness from its maximal value. Dependence of the robust fitted phase on the ratio of functional to nonfunctional parts has been analyzed in depth. As the former ratio is increased, the selection pressure to achieve this phase is drastically increased, whereas, if the phase is achieved, it can persist for slightly higher noise. This suggests the relevance of having sufficient nonfunctional parts in biological systems. In addition, correlation between robustnesses to noise and to mutation is formulated as a proportionality between susceptibilities to external field and to coupling change.

II. DOUBLE REPLICA THEORY

Following [15,16] we study the evolution of the relationship between phenotype and genotype by representing phenotypes as spin configurations, and genotypes as interaction matrix for spins. In a system of N spins, each spin i can take values $s_i \in \{-1, 1\}$ and is linked to exactly $N-1$ other spins, thus forming a fully connected network. Here the evolution progresses through the change in genotype, whereas the selection is based on the phenotype, resulting in an evolution of the genotype-phenotype relation. Moreover, fitness is determined by a subset of target spins denoted by $\mathcal{T}$. Those spins that do not contribute to the fitness are called nontarget. In general, the fitness Ψ is some field that acts on J_{ij} but whose value depends only on s_i, $i \in \mathcal{T}$. How such dependence is explicitly described is model specific and will not limit the use of our approach. See Eq. (A1) in Appendix A for an example of Ψ given by the target spin configurations at equilibrium [15,16].

Stochastic dynamics of phenotypes are considered as the evolution of spin configurations at a temperature T_s according to a Hamiltonian $H_S = -\sum_{i<j} J_{ij} s_i s_j$ [24]. Here the couplings J_{ij} are regarded as *fixed* over the course of the spin evolution that follows a Glauber update because they are assumed to evolve on much slower timescale than that of the spins. Furthermore, the couplings are symmetric, i.e., $J_{ij} = J_{ji}$, and, *initially*, are independently and identically distributed by a Gaussian distribution with zero mean and the variance $J^2 := \mathrm{var}(J_{ij}) = N^{-1}$. The coupling matrix $\mathbf{J}$ includes interactions between target spins $J_{ij}^{(tt)}$ for $i \in \mathcal{T}$ and $j \in \mathcal{T}$, those between nontarget spins $J_{ij}^{(oo)}$ for $i \in \mathcal{T}$ and $j \in \mathcal{T}$, and those between target spin and nontarget spin $J_{ij}^{(to)}$ for $i \in \mathcal{T}$ and $j \in \mathcal{T}$. Let $S_\mathcal{T}$ and $S_\mathcal{O}$ denote the subsystems of target spins (with their interactions $\mathbf{J}^{(tt)}$) and the subsystem of nontarget spins (with the couplings $\mathbf{J}^{(oo)}$ among them), respectively. The Hamiltonian of the full system denoted by S can be decomposed into

$$H_S = \underbrace{-\sum_{i<j\in\mathcal{T}} J_{ij}^{(tt)} s_i s_j}_{H_\mathcal{T}} \underbrace{-\sum_{i<j\notin\mathcal{T}} J_{ij}^{(oo)} s_i s_j}_{H_\mathcal{O}} \underbrace{-\sum_{\substack{i\in\mathcal{T}\\ j\notin\mathcal{T}}} J_{ij}^{(to)} s_i s_j}_{H_{\mathcal{T}\mathcal{O}}}, \quad (1)$$

where $H_\mathcal{T}$ and $H_\mathcal{O}$ are the Hamiltonians of the subsystems $S_\mathcal{T}$ and $S_\mathcal{O}$, respectively, while $H_{\mathcal{T}\mathcal{O}}$ describes the interactions between these subsystems.

Now we introduce the effective potential to obtain the distribution of $\mathbf{J}$ [25]. For it, we consider a continuous Langevin-type dynamics for the couplings,

$$\frac{dJ_{ij}}{d\tau} = -\frac{1}{N}\frac{\partial V}{\partial J_{ij}} + \frac{1}{\sqrt{N}}\xi_{ij}(\tau), \quad (2)$$

where $V = V(\mathbf{J})$ is the potential of all the couplings and ξ_{ij} is the white noise whose intensity equal to the temperature T_J. The factors $1/N$ and $1/\sqrt{N}$ in front of the potential and the noise term, respectively, ensure a correct relationship between the drift and diffusive parts of the Langevin equation. If the couplings were independent of each other, the potential would simply take the form of the potential of a free Brownian

particle,

$$V_0 = \frac{N}{2}\sum_{i<j} J_{ij}^2. \tag{3}$$

However, in the presence of fitness we need an additional term. Here we assume that the fitness would be maximized if a global alignment is established among target spins, so that we introduce

$$\Psi = \frac{1}{N_t}\left|\sum_{i\in\mathcal{T}} s_i\right|, \tag{4}$$

where N_t is the size of $\mathcal{T}$. Under this fitness that favors the alignment of target spins, the couplings are necessarily subjected to a fitness field K [26]:

$$K = \frac{1}{\beta_J}\frac{\partial}{\partial J_{ij}} \ln\left(\sum_{\{s_i\}} \exp\left\{\frac{\beta_J}{N_t}\sum_{i<j\in\mathcal{T}} J_{ij}\left|\sum_{i\in\mathcal{T}} s_i\right|\right\}\right) \tag{5}$$

or equivalently, V needs to be modified from V_0 into

$$V = V_0 - \frac{1}{\beta_J}\ln\left(\sum_{\{s_i\}} \exp\left\{\frac{\beta_J}{N_t}\sum_{i<j\in\mathcal{T}} J_{ij}\left|\sum_{i\in\mathcal{T}} s_i\right|\right\}\right).$$

Without the evolution of genotypes, the Hamiltonians $H_{\mathcal{T}}$, $H_{\mathcal{O}}$, and $H_{\mathcal{TO}}$ dictate the spins to adapt to a set of fixed couplings $\mathbf{J}$ in order to minimize each term of Eq. (1) through the spin dynamics. Such adaptation results in an accordance between the state of $J_{ij}^{(tt)}$ and $s_i s_j$ for $i, j \in \mathcal{T}$, that between the state of $J_{ij}^{(oo)}$ and $s_i s_j$ for $i, j \notin \mathcal{T}$, and that between the state of $J_{ij}^{(to)}$ and $s_i s_j$ for $i \in \mathcal{T}$, $j \notin \mathcal{T}$. As long as this kind of accordance exits, it is insufficient to consider the evolving couplings with selection force given in Eq. (5) only. A link between two spins hence necessarily needs to adapt to the joint state of these spins. Due to the timescale separation between the phenotype and the genotype dynamics, the direction of change of genotypes is determined by the equilibrium correlation of the phenotypes, and since J_{ij} is symmetric, it needs to be

$$dJ_{ij}/d\tau \propto \langle s_i s_j\rangle_{T_s}.$$

This is equivalent to having a potential of the form [27]

$$V_a = V - \frac{1}{\beta_s}\ln\left(\sum_{\{s_i\}} \exp\left\{\beta_s \sum_{i<j} J_{ij} s_i s_j\right\}\right). \tag{6}$$

The stochastic process induced by Eq. (2) under this potential admits an equilibriumlike stationary *joint* distribution $\mathbb{P}(\mathbf{J}^{(tt)}, \mathbf{J}^{(oo)}, \mathbf{J}^{(to)})$ of Boltzmann form (with associated temperature T_J), i.e.,

$$\mathbb{P}(\mathbf{J}^{(tt)}, \mathbf{J}^{(oo)}, \mathbf{J}^{(to)}) = e^{-\beta_J V_a}/\mathcal{Z}_{\text{total}},$$

where $\mathcal{Z}_{\text{total}} = \sum_{\{\mathbf{J}\}} e^{-\beta_J V_a}$. Instead of calculating this distribution, we introduce our approximate approach, in which $J_{ij}^{(to)}$ are assumed to always attain equilibrium well before $J_{ij}^{(tt)}$ and $J_{ij}^{(oo)}$ and hence can be adiabatically eliminated. As a consequence, only the weights of equilibrium configurations of $\mathbf{J}^{(to)}$ contribute to the stationary distributions:

$$\mathbb{P}_{\mathcal{T}}(\mathbf{J}^{(tt)}) = \lim_{\tau\to\infty} \mathbb{P}_{\mathcal{T}}(\mathbf{J}^{(tt)}, \tau),$$
$$\mathbb{P}_{\mathcal{O}}(\mathbf{J}^{(oo)}) = \lim_{\tau\to\infty} \mathbb{P}_{\mathcal{O}}(\mathbf{J}^{(oo)}, \tau),$$

where $\mathbb{P}_{\mathcal{T}}(\mathbf{J}^{(tt)}, \tau)$ and $\mathbb{P}_{\mathcal{O}}(\mathbf{J}^{(oo)})$ are the time-dependent solutions of the corresponding Fokker-Planck equations with the effective potentials V_{tt} for $\mathbf{J}^{(tt)}$ and V_{oo} for and $\mathbf{J}^{(oo)}$, respectively [28]. To obtain the distribution only of $\mathbf{J}^{(tt)}$ and $\mathbf{J}^{(oo)}$, we first replace $\mathbf{J}^{(to)}$ as the given matrix by that obtained self-consistently from the equilibrium distribution. For it we need to modify V_a in such a way that the influence of $\mathbf{J}^{(to)}$ on $\mathbf{J}^{(tt)}$ ($\mathbf{J}^{(oo)}$) can be taken into account in the effective potential V_{tt} (V_{oo}). The *joint* effect of $H_{\mathcal{TO}}$ and $H_{\mathcal{T}}$ on the dynamics of target spins suggests that the dependence of $J_{ij}^{(tt)}$ and $J_{ik}^{(to)}$ on each other arises from any triad formed between $(i, j) \in \mathcal{T}$ and $k \notin \mathcal{T}$, i.e., via $J_{ij}^{(tt)} J_{ik}^{(to)} J_{jk}^{(to)}$. This influence is represented by the frustration [17], which implies that an optimal spin configuration (s_i^*, s_j^*, s_k^*) can only be established if the relation $J_{ij}^{(tt)} J_{ik}^{(to)} J_{jk}^{(to)} > 0$ holds [optimality in this context means that $H_{\mathcal{TO}}$ and $H_{\mathcal{T}}$ can be lowered simultaneously by (s_i^*, s_j^*, s_k^*)]. Since for any given pair of $(i, j) \in \mathcal{T}$ there are $N - N_t$ triads Δ_k formed between it and a nontarget spin $k \notin \mathcal{T}$, the total effect of frustration is given by

$$F_{ij}^{(t\text{-}o\text{-}t)} = (N - N_t)^{-1}\sum_{k=1}^{N-N_t} J_{ik}^{(to)} J_{jk}^{(to)}. \tag{7}$$

Likewise, frustration among all N_t triads $\tilde{\Delta}_k$ formed between a given pair of nontarget spins $i \notin \mathcal{T}$ and $j \notin \mathcal{T}$ with $k \in \mathcal{T}$ induces a force $F_{(o\text{-}t\text{-}o)}$ on the state of $J_{ij}^{(oo)}$:

$$F_{ij}^{(o-t-o)} = N_t^{-1}\sum_{k=1}^{N_t} J_{ik}^{(to)} J_{jk}^{(to)}. \tag{8}$$

The proposed scheme just allows us to define the effective potential V_{tt} for $\mathbf{J}^{(tt)}$ and V_{oo} for and $\mathbf{J}^{(oo)}$ as

$$V_{tt} = V_a - \frac{1}{\beta_J}\ln\left(\sum_{\{J_{ij}^{(to)}\}} \exp\left\{\beta_J \sum_{i<j\in\mathcal{T}} J_{ij}^{(tt)} F_{ij}^{(t-o-t)}\right\}\right),$$

$$V_{oo} = V_a - \frac{1}{\beta_J}\ln\left(\sum_{\{J_{ij}^{(to)}\}} \exp\left\{\beta_J \sum_{i<j\notin\mathcal{T}} J_{ij}^{(oo)} F_{ij}^{(o-t-o)}\right\}\right).$$

The stationary distributions induced by the diffusion process in Eq. (2) with these effective potentials have a Boltzmann form, $\mathbb{P}_{\mathcal{T}}(\mathbf{J}^{(tt)}) = e^{-\beta_J V_{tt}}/\mathcal{Z}_{\mathcal{T}}$ and $\mathbb{P}_{\mathcal{O}}(\mathbf{J}^{(oo)}) = e^{-\beta_J V_{oo}}/\mathcal{Z}_{\mathcal{O}}$, where $\mathcal{Z}_{\mathcal{T}}$ and $\mathcal{Z}_{\mathcal{O}}$ are the partition function of the genotypes $\mathbf{J}^{(tt)}$ and that of the genotypes $\mathbf{J}^{(oo)}$, respectively [29]. Here we replace $\mathbf{J}^{(to)}$ by the replica matrix $\sigma_i^k := J_{ik}^{(to)}$, to be obtained. (This stepwise scheme is valid, as we are concerned with the equilibrium property). Denoting $n := T_s/T_J$, we can compute $\mathcal{Z}_{\mathcal{T}}$ and $\mathcal{Z}_{\mathcal{O}}$ as

$$\mathcal{Z}_{\mathcal{T}} = \int \prod_{i<j\in\mathcal{T}} dJ_{ij}^{(tt)} \sum_{\left\{s_i;s_i^a;\sigma_i^k\right\}_{i\in\mathcal{T}}} \exp\left\{\beta_J \sum_{i<j}\left[-\frac{N_t}{2}\left(J_{ij}^{(tt)}\right)^2 + J_{ij}^{(tt)}\left(\frac{1}{N_t}\left|\sum_{i\in\mathcal{T}} s_i\right| + \underbrace{\frac{1}{n}\sum_{a=1}^{n} s_i^a s_j^a}_{s\text{ replicas}} + \underbrace{\frac{1}{N-N_t}\sum_{k=1}^{N-N_t}\sigma_i^k\sigma_j^k}_{\sigma\text{ replicas}}\right)\right]\right\},$$

$$\mathcal{Z}_{\mathcal{O}} = \int \prod_{i<j\notin\mathcal{T}} dJ_{ij}^{(oo)} \sum_{\left\{s_i^a;\sigma_i^k\right\}_{i\notin\mathcal{T}}} \exp\left\{\beta_J \sum_{i<j}\left[-\frac{N-N_t}{2}\left(J_{ij}^{(oo)}\right)^2 + J_{ij}^{(oo)}\left(\tilde{K} + \underbrace{\frac{1}{n}\sum_{a=1}^{n} s_i^a s_j^a}_{s\text{ replicas}} + \underbrace{\frac{1}{N_t}\sum_{k=1}^{N_t}\sigma_i^k\sigma_j^k}_{\sigma\text{ replicas}}\right)\right]\right\}.$$

In writing these equations, we assume that the fitness acts only on $J_{ij}^{(tt)}$ [30] and hence in $\mathcal{Z}_{\mathcal{O}}$ we replace the fitness field K by a constant $\tilde{K}$, which eventually will be set to 0 by virtue of calculations of the observables for nontarget spins. Although neglecting the fitness's effect on $J_{ij}^{(oo)}$ does not follow exactly the above-mentioned implementation of the model, we expect that this holds true in the long-time limit because otherwise both target and nontarget configurations at equilibrium would determine the fitness. This restriction hence corresponds to a first-order approximation of the fitness's effect, while a term that affects the dynamics of $J_{ij}^{(oo)}$ and $J_{ij}^{(to)}$ is considered to be of higher order.

We here propose to interpret σ_i^k as the kth replica of another variable $\sigma_i \in \{-1, 1\}$ that is also located at the site i of the graph (generally $\sigma_i \neq s_i$). To distinguish these different types of replicas from each other, we call s_i^a spin replicas and σ_i^k coupling replicas. Following this interpretation, apart from n that appears as the number of spin replicas s_i^a, $a = \{1, \ldots, n\}$, in $\mathcal{Z}_{\mathcal{T}}$ and $\mathcal{Z}_{\mathcal{O}}$ [31], we thus have $N - N_t$ coupling replicas, σ_i^k, for $i \in \mathcal{T}$ and $k = \{1, \ldots, N - N_t\}$, and N_t coupling-replicas, σ_i^k, for $i \notin \mathcal{T}$ and $k = \{1, \ldots, N_t\}$ respectively. As, in general, none of these numbers are zero, our double-replica approach does not correspond to the conventional quenched limit in spin-glass models [17]. Once setting $K = 0$ and neglecting the terms corresponding to $F^{(t\text{-}o\text{-}t)}$ and $F^{(o\text{-}t\text{-}o)}$, we recover the model of Coolen *et al.* for neural systems with dynamic synapses [19]. In contrast to the use of a Hamiltonian for the couplings adopted in [23], here we have introduced the effective potential for couplings that, by using the timescale separation between the dynamics of genotypes and that of phenotypes, allows for the integration of the spin dynamics specified by the Hamiltonian (1) into the Langevin dynamics of the couplings through the second term in Eq. (6).

Here we characterize the equilibrium behavior of the model by the average fitness m, the overlap between different spin replicas a and b, q_{ab}, and the correlation between adjacent links Q. Additionally, we want to quantify the mean value of the couplings among the target spins only Φ. Let $\mathbb{E}[\cdot]$ and $\tilde{\mathbb{E}}[\cdot]$ denote ensemble averages over $\mathbb{P}_{\mathcal{T}}(\mathbf{J}^{(tt)})$ and $\mathbb{P}_{\mathcal{O}}(\mathbf{J}^{(oo)})$, respectively. These order parameters are given by

$$m_a = \mathbb{E}\left[s_i^a\right]_{i\in\mathcal{T}}, \quad q_{ab} = \mathbb{E}\left[s_i^a s_i^b\right]_{i\in\mathcal{T}}, \tag{9a}$$

$$Q_{kk'} = \mathbb{E}\left[J_{ik}^{(to)} J_{ik'}^{(to)}\right]_{\substack{i\in\mathcal{T}\\ k,k'\notin\mathcal{T}}}, \tag{9b}$$

$$\Phi = \mathbb{E}\left[J_{ij}^{(tt)}\right]_{i,j\in\mathcal{T}}. \tag{9c}$$

Similarly, for the nontarget spins we have

$$\tilde{m}_a = \tilde{\mathbb{E}}\left[s_i^a\right]_{i\notin\mathcal{T}}, \quad \tilde{q}_{ab} = \tilde{\mathbb{E}}\left[s_i^a s_i^b\right]_{i\notin\mathcal{T}}, \tag{10a}$$

$$\tilde{Q}_{kk'} = \tilde{\mathbb{E}}\left[J_{ik}^{(to)} J_{ik'}^{(to)}\right]_{\substack{i\notin\mathcal{T}\\ k,k'\in\mathcal{T}}}, \tag{10b}$$

$$\tilde{\Phi} = \tilde{\mathbb{E}}\left[J_{ik}^{(oo)}\right]_{i,j\notin\mathcal{T}}. \tag{10c}$$

In the thermodynamics limit, $N \to \infty$ and $N_t \to \infty$, while keeping $p = N_t/N$ fixed, using a replica symmetric ansatz for the variables $m_a = m$, $q_{ab} = q$ and $Q_{kk'} = Q$, $\tilde{m}_a = \tilde{m}$; $\tilde{q}_{ab} = \tilde{q}$ and $\tilde{Q}_{kk'} = \tilde{Q}$, as well as $M_{ak} = M$ and $\tilde{M}_{ak} = \tilde{M}$, where $M_{ak} = \mathbb{E}[s_i^a J_{ik}^{(to)}]_{\substack{i\in\mathcal{T}\\ k\notin\mathcal{T}}}$ and $\tilde{M}_{ak} = \tilde{\mathbb{E}}[s_i^a J_{ik}^{(to)}]_{\substack{i\notin\mathcal{T}\\ k\in\mathcal{T}}}$, we obtain the following free energy densities:

$$f_{\mathcal{T}}^{\mathrm{RS}} = \frac{1}{2}\left\{\frac{q}{n} + \frac{Q}{N-N_t} + \frac{(n-1)q^2}{2n} + \frac{Q^2}{2} + M^2\right\} - \frac{1}{\beta_J}\ln\left[\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k} I_{kz}\right] \tag{11a}$$

$$f_{\mathcal{O}}^{\mathrm{RS}} = \frac{1}{2}\left\{\frac{\tilde{q}}{n} + \frac{\tilde{Q}}{N_t} + \frac{(n-1)\tilde{q}^2}{2n} + \frac{\tilde{Q}^2}{2} + \tilde{M}^2\right\} - \frac{1}{\beta_J}\ln\left\{\sum_{k=0}^{N_t}\binom{N_t}{k}\tilde{I}_k\right\}. \tag{11b}$$

From the extremum condition of these free energies we can compute all the model order parameters via a set of self-consistency equations. These equations as well as the functions I_{kz} and $\tilde{I}_k$ are given in Appendix B [32]. The use of the replica symmetry is justified in most of the (T_s, T_J) parameter space from the stability analysis [33]. At low T_s and T_J the replica symmetry is broken, which we will not explore fully. Nevertheless we will discuss later how robustness of phenotypes, postulated for biological systems that reproduce similar offspring, is lost in that scenario. The replica-symmetric free energy densities allow us to derive

$$\Phi = \frac{1}{N_t}[\Psi + m^2 + r^2], \tag{12a}$$

$$\tilde{\Phi} = \frac{\tilde{m}^2 + \tilde{r}^2}{N - N_t}, \tag{12b}$$

where r and $\tilde{r}$ are defined and computed in Appendix B.

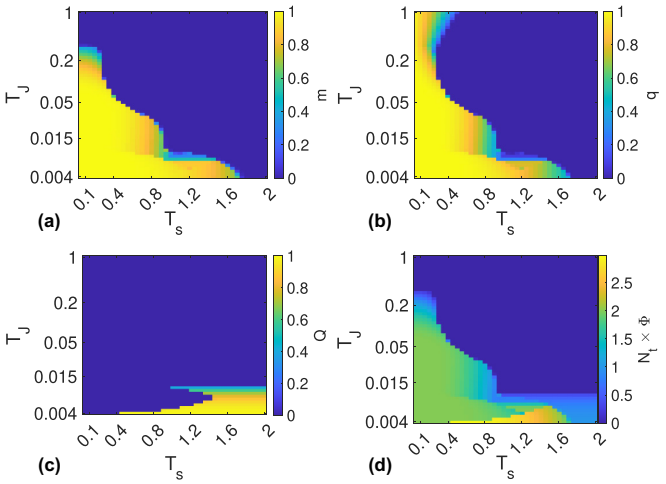

FIG. 1. Magnetization for target spins m (a). Overlap between different replicas for target spins q (b). Averaged correlation of a pair of couplings between a target and a nontarget spin that share a common nontarget spin Q (c). Averaged frustration among target spins Φ (d). Here $N_t = 10$, $N = 100$. Note the y axis is on logarithmic scale.

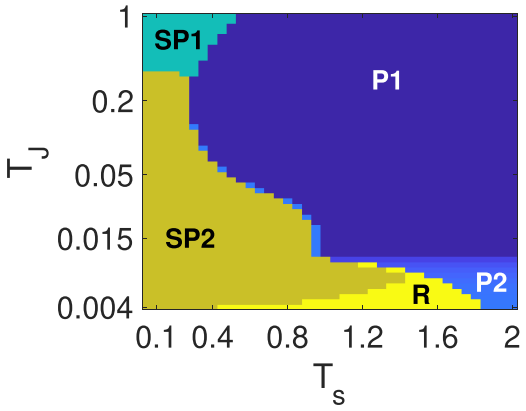

FIG. 2. The model phase diagram. Here **SP1** denotes the spin-glass phase with $m = \tilde{m} = Q = \tilde{Q} = 0$ and $q, \tilde{q} > 0$; **SP2** denotes the spin-glass phase with $\tilde{m} = Q = 0$ and $q, m > 0$ (within this region, both $\tilde{q}$ and $\tilde{Q}$ can be either zero or non-zero, see Appendix D); **P1** denotes the paramagnet phase with $m = \tilde{m} = q = \tilde{q} = Q = \tilde{Q} = 0$; **P2** denotes the paramagnet phase with $m = \tilde{m} = q = \tilde{q} = \tilde{Q} = 0$ but $Q > 0$; **R** denotes the robust fitted phase with $\tilde{m} = \tilde{q} = \tilde{Q} = 0$ but $m, q, Q > 0$. Here $N_t = 10$, $N = 100$. Note the y-axis is on logarithmic scale.

III. PHASE DIAGRAM

In Fig. 1 we depict the order parameters as function of the temperature T_s and T_J for a particular choice of $N = 100$ and $N_t = 10$. Here for each point (T_s, T_J) of the phase diagram, we solve numerically the set of mean-field equations for m, q, Q, while computing Φ from the knowledge of these quantities. In terms of only the magnetization m and the overlap between spin-replicas q for target spins, we observe three distinct phases that are typical for spin-glass systems, namely, $m = q = 0$ (paramagnet phase), $m = 0, q > 0$ (spin-glass phase), and $m > 0, q > 0$ with $\sqrt{q} > m$ (target-ferromagnet phase, "t-ferro" in short). The transitions between the phases are second order at small T_s/T_J, but become discontinuous (first order) at large T_s/T_J. At a much lower value of T_J there is a region where, apart from having a nonzero magnetization of target spins, the order parameter $Q = \langle J^{(to)} J^{(to)} \rangle$ starts to become nonzero. As can be anticipated from Eq. (12a), the mean value of $J_{ij}^{(tt)}$ also varies from region to region in accordance with the change of m and that of Q. Note that in sharp contrast to the transition between paramagnet and spin glass, which is similar to that of the Sherrington-Kirkpatrick (SK) model, the phenotype-genotype coupling results in a repositioning of the boundary between spin glass and t-ferro. Such difference arises from the nonzero correlation of the genotypes. Expanding the free energy $f_{\mathcal{T}}^{\mathrm{RS}}$ for small m and q, the transition between paramagnet and t-ferro occurs at $T_s^{\mathbf{P}\to\mathbf{F}} = \kappa$, where $\kappa = 2^{-N_t} \sum_{z=0}^{N_t} \binom{N_t}{z} |N_t - 2z|/N_t$, while the spin-glass to t-ferro transition occurs at $[1 + (n-1)q(\beta_s^{\mathbf{SP}\to\mathbf{F}})](\beta_s^{\mathbf{SP}\to\mathbf{F}}\kappa) = 1$. We also check that both the magnetization $\tilde{m}$ of the nontarget spins and the average value $\tilde{\Phi}$ of $J_{ij}^{(oo)}$ are always zero, as the nontarget spin subsystem remains frustrated all the time, while the spin overlap $\tilde{q}$ can undergo a transition from paramagnet to spin glass, in the same way as the SK model. The phase diagrams of these quantities are given in in Appendix D. Combining the behavior of the order parameters altogether, we obtain the model phase structure in Fig. 2. It contains five distinct regions. At low genotypic selection pressure $T_J \geqslant e^{-1}$, only the first spin-glass **SP1** and the paramagnet **P1** phases with zero fitness are observed. However, as the genotypic selection pressure increases other phases emerge. At sufficiently low T_J, a robust fitted phase denoted by **R** ($m, q, Q > 0$) emerges in an intermediate range of T_s (here $\tilde{m} = \tilde{q} = \tilde{Q} = 0$). Adjacent to this phase on the side of high phenotypic noise is the second paramagnet phase **P2** where the fitness value is low ($m = q = 0$) but there exists some structure in the genotypes such that $Q > 0$. On the other hand, for lower T_s, the system is in the second spin-glass phase **SP2** with high fitness but nonrobust genoptypes ($m \simeq q \simeq 1, Q = 0$). In particular, the transition from **R** to **SP2** is marked by a replica symmetry breaking (RSB) which indicates the loss of stability of the replica symmetric (RS) solutions [34]. The broad distribution of gene-gene correlations in the RSB phase implies that the genotype of offsprings is not preserved, in contrast to the RS phase. In the biological context, this means that replication is no longer stable so that genotypes are not conserved over generations.

Overall, the phase diagram agrees with what was observed numerically in [15,16]. However, thanks to the explicit account of the coupling replicas, so that Q can be treated as an order parameter upon which the free energy density depends, we discover the existence of the second paramagnet phase **P2** that was not reported before. This phase can be interpreted as a *precursor* region, in which genotypes are structured in such a way that supports ferromagnetic ordering among target spins, and hence have potentiality to acquire a high fitness, but due to the high fluctuation induced by T_s, this fitness cannot be maintained. Furthermore, by considering separately the effective dynamics of the target and nontarget subsystems, $S_{\mathcal{T}}$ and $S_{\mathcal{O}}$, our approach can differentiate the phase **SP1** from **SP2**. The previous approach [23] only stressed the distinct arrangement of target spins in the **SP2** region,

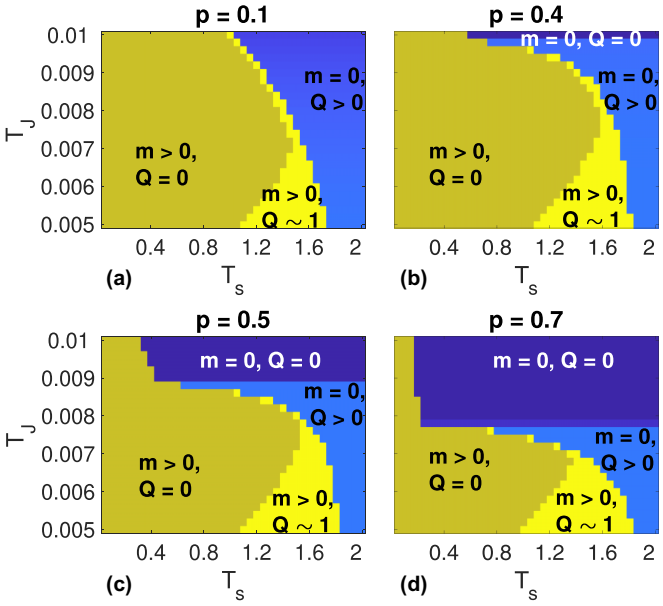

FIG. 3. The phase diagrams obtained by combining the behavior of m as function of T_s and T_J with that of Q for different values of p, at sufficiently low T_J. Here $N = 100$.

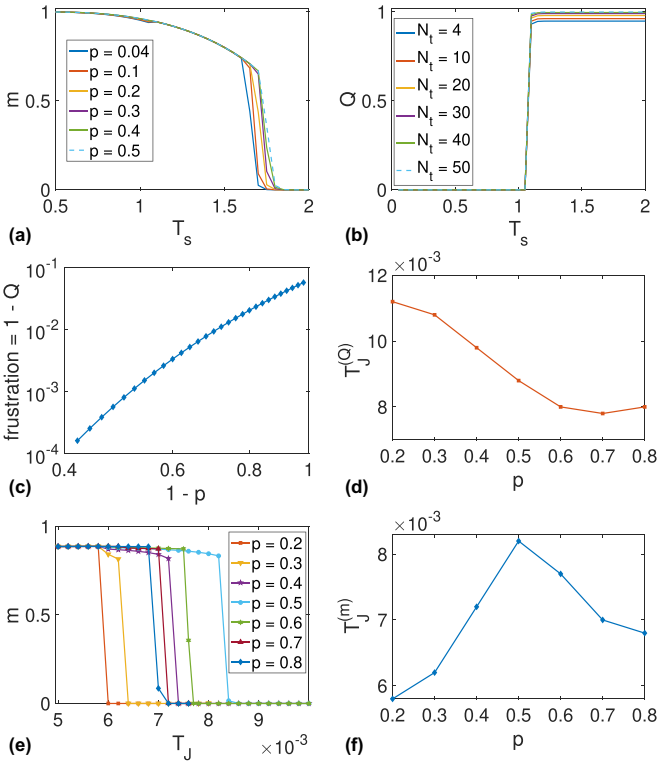

FIG. 4. (a) Magnetization of target spins m as function of T_s at fixed $T_J = 0.005$ for various numbers of target spins, $p = 0.04, 0.1, 0.2, 0.3, 0.4, 0.5$. (b) The same for genetic overlap Q. (c) Frustration defined as $1 - Q$ as a function of the number of nontarget spins in the robust fitted phase (i.e., $T_s \in [T_c^{(1)}, T_c^{(2)}]$) at fixed $T_J = 0.005$. (d) The highest value of T_J at which Q remains nonzero as a function of p. (e) Magnetization of target spins m as a function of T_J at fixed $T_s = 1.3$ for various fractions of target spins, $p = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$. (f) The value of T_J at which the magnetization of target spins, m, drops to zero at fixed $T_s = 1.3$ for different p from (e). Behavior similar to (e) and (f) is observed for other $T_s \in [T_c^{(1)}, T_c^{(2)}]$. Here $N = 100$.

where the subsystem of target spins becomes ferromagnetic whereas that of nontarget ones remains spin glass. Our present approach shows that this is no longer true for a high value of T_J. Upon increasing T_J, this ferromagnetic ordering is destroyed by genotypic fluctuations. The same structure of the phase diagrams is also observed for $N = 50$, $N_t = 5$ and $N = 200$, $N_t = 20$; see Appendix C for the phase diagram obtained in these cases, where β_J is rescaled according to N (and N_t). Accordingly, it is shown that the selection pressure needed to achieve robustness increases with the number of spins N at a fixed fraction N_t/N. In addition, the present analysis allows one to obtain quantitative dependence of genotypic and phenotypic robustness on the fraction of targets. The phase diagram in Fig. 2 includes global information of the system including weak selection region without achieving nonzero fitness, m, of target, whereas there is biological interest if the fitted state is evolved robustly by the selection. To this end we focus on the low-T_J region of the phase diagram to explore the dependence of the system behavior on the fraction $p = N_t/N$ of target spins. While overall the phase structure is similar for different p in Fig. 3, in particular the robust fitted (yellow) region seems to change slightly with increasing p; the relative size and exact location of all the other phases vary with p. This suggests that a more quantitative analysis is needed to understand the genotype-phenotype relationship as a function of p. We carry on this analysis in the next section.

IV. STRUCTURE OF THE ROBUST FITTED PHASE

The most relevant region in the phase diagram is the robust fitted phase **R**, which is characterized by both the high fitness ($m > 0$) and robustness ($Q > 0$). For a sufficiently low given T_J (i.e., high selection pressure), the phase is bounded by $T_s \in [T_c^{(1)}, T_c^{(2)}]$. Below $T_c^{(1)}$, Q goes to zero, and above $T_c^{(2)}$, m goes to zero, whereas these transition points depend on T_J. We first investigate the dependence on p of the **R** region by fixing T_J. In Figs. 4(a) and 4(b) we fixed $T_J = 0.005$. First, for a wide range of $p \in [0.04, 0.5]$, the temperature $T_c^{(1)}$ of the latter transition in (b) does not depend on p (see Appendix D for the zoom-in of a small dip of m at this point). On the other hand, $T_c^{(2)}$ slightly increases with increasing p in (a), indicating that the fitness of the high p case is relatively more robust to noise than the low p case.

In Fig. 4(c) $1 - Q$ is almost constant against T_s in the **R** phase. This constant value was found to increase with the number of nontarget spins. Note that $Q \sim 1$ implies that the offspring of genotypes are preserved. The increase in $1 - Q$, thus means the increase in redundancy of genotypes, as is supported by a larger number of non-targets. Such a redundancy has another meaning in the context of spin-glass systems, where it is indeed equal to the local frustration [in the (t-o-t) triads with positive $J^{(tt)}$].

In contrast, apart from the two critical points $T_c^{(1)}$ and $T_c^{(2)}$, the fitness m does not depend on p. It follows a unique curve, independent of p. Even though the increase in genetic

heterogeneity $1-Q$ for more nontargets may perturb the target spin configuration, the fitness m remains unchanged even for smaller p.

Then, we estimate the critical value of T_J below which the **R** phase can exist. While this critical value denoted by $T_J^{(Q)}$ can depend on T_s, as seen in Figs. 2 and 3, it can be approximately identified with the upper part of the **P2** phase from the phase diagram. In Fig. 4(d) $T_J^{(Q)}$ is shown to decrease with p. This result together with that in Fig. 4(a) mean that the higher p is, the higher the selection pressure that is needed to achieve robustness, but once it is achieved, a system with larger p is more robust with respect to phenotypic noise than one with smaller p.

Finally, we examine the fitness as function of T_J at fixed $T_s = 1.3$ for various p in Fig. 4(e). While fitness decreases with T_J, its behavior with the increase in p is nonmonotonic. This behavior is further shown in Fig. 4(f) where the critical genotypic noise $T_J^{(m)}$ at which the fitness becomes nonzero is plotted versus p. Similar behavior is observed for other $T_s \in [T_c^{(1)}, T_c^{(2)}]$. The result supports $p = 0.5$ as the maximal value of $T_J^{(m)}$, suggesting the existence of an optimal fraction of target spins to acquire high fitness in this intermediate range of T_s [35].

V. MUTATIONAL SUSCEPTIBILITY AND PHENOTYPIC SUSCEPTIBILITY IN THE ROBUST FITTED PHASE

Correlation between variances of phenotypes due to genetic changes and to noise has been discussed in both experiments and simulations, and relationships to robustness have been discussed both theoretically [10,14,36–39] and experimentally [40–43]. In statistical physics, this issue can be analyzed in terms of susceptibility, as it is proportional to the variance. Then, we need to study the susceptibility due to genetic mutation, in addition to the standard susceptibility.

In the context of this model, mutations are defined as those change of the genotypes J_{ij} that might happen spontaneously and independently of the dynamics specified previously. Let $\delta\Psi_i(\delta J_{jk})$ denote the change of the average local magnetization of a target spin i [44] upon mutating a genotype $J_{jk} \to J_{jk} + \delta J_{jk}$. The mutational susceptibility of this target spin with respect to such a change $M_{i,jk}$ then can be defined as

$$M_{i,jk} = \lim_{\delta J_{jk}\to 0} \delta\Psi_i(\delta J_{jk})/\delta J_{jk} = \langle s_i s_j s_k\rangle - \langle s_i\rangle\langle s_j s_k\rangle.$$

In general, $J_{jk} \in \mathbf{J} = \mathbf{J}^{(tt)} \cup \mathbf{J}^{(oo)} \cup \mathbf{J}^{(to)}$. However, since fitness is determined solely by the configurations of target spins at equilibrium, we consider only $J_{jk} \in \mathbf{J}^{(tt)}$ and show that the average of this mutational susceptibility over all triples $(i, j, k) \in \mathcal{T}$ is equal to

$$M := \frac{1}{\binom{N_t}{3}} \sum_{(i,j,k)} M_{i,jk} = 2\beta_J^2 m\chi_m - \beta_J \frac{\partial^3 f}{\partial h^3}\Big|_{h=0} \tag{13}$$

where $\chi_m := -\lim_{h\to 0} \partial^2 f/\partial h^2$ is the susceptibility of target spins. The quantities M and χ_m correspond to the susceptibility to mutation of the genotypes and susceptibility to perturbation by an external field, h, respectively. We can expect that in the robust fitted phase there exists a relation between M and χ_m [45]. In fact, let $X := \lim_{h\to 0} \partial^3 f_{\mathcal{T}}/\partial h^3$. For L given in Eq. (B3a) in Appendix B (L/β_J has the meaning of an effective Hamiltonian that is defined in the combined space $\{s_i^a, \sigma_i^k\}$ of s replicas and J replicas), according to the definitions

$$X \propto \sum_{a,b,c=1}^{n} \mathrm{Tr}(s^a s^b s^c e^L)/\mathrm{Tr}(e^L),$$

$$m := \langle s^a\rangle = \frac{1}{n}\sum_{a=1}^{n} \mathrm{Tr}(s^a e^L)/\mathrm{Tr}(e^L),$$

$$\chi_m := \lim_{h\to 0}\frac{\partial m}{\partial h} = \frac{\beta_s}{n}\sum_{a,b=1}^{n} \mathrm{Tr}(s^a s^b e^L)/\mathrm{Tr}(e^L),$$

the symmetry between different replicas in the robust fitted phase implies that the third moment X^{RS} is proportional to the product of the first and second moments, $m^{\mathrm{RS}}\chi_m^{\mathrm{RS}}$. Therefore, approximately, $M \propto \chi_m^{\mathrm{RS}}$. This proportionality between the two susceptibilities, implying a correlation between phenotypic changes due to genetic variation and those in response to environmental perturbations [46], does not exist in the RSB phase, as the second term in Eq. (13) is no longer proportional to χ_m.

VI. DISCUSSION

In the paper we propose an approach towards biological evolution due to the interrelationship between genotype and phenotype where fitness is determined solely by the latter but not by the former. Though the emergence of structured genotypes from initially random couplings under this relation has been numerically reported, apart from a study which imposed a specific condition on the couplings [23], this has not been studied analytically yet. We here are able to tackle this problem thanks to what we termed *double-replica* theory. Within this framework we obtain the phase diagram, that is classified not only by the fitness but also by the overlap in dual replicas. The diagram is not only in good agreement with previous studies (including paramagnet, t-ferro, and robust fitted phases, all existing at sufficiently low T_J), but also contains additional phases. These include the first spin-glass phase **SP1** and the second paramagnet phase **P2**. The former corresponds to a system with both target and nontarget spins residing in a spin-glass phase (at low selection pressure), while the latter corresponds to a paramagnetic phase for all spins but retains genetic correlations encoding target and nontarget couplings (at high selection pressure and high T_s). Here, even though the genotypes favor a high value of fitness, due to large fluctuations induced by T_s, such a value can not be maintained over generations. The existence of the phase suggests that even though the average fitness is zero due to large noise, there exists a genetic precursor to generate individuals with nonzero fitness. The relevance of this scenario to evolutionary biology needs to be explored in future, though.

The system can only acquire high fitness at some $T_s \leqslant T_c^{(2)}$, where the fitness increases discontinuously. If T_s is too low,

then RSB will happen, leading to a phase without *genetic* overlap, where biologically required robustness of genotypes is lost. Hence a lower bound of $T_s \geqslant T_c^{(1)}$ is necessary to have RS and robustness, accordingly.

From this approach, the target-fraction dependence of genotypic and phenotypic robustness can also be understood quantitatively. Such dependence is quantified via the behavior of the fitness m and genetic redundancy $1-Q$ in the robust region bounded by $T_c^{(1)}$ and $T_c^{(2)}$. Here we find that a genetically homogeneous population can only be robustly reproduced under a sufficiently high selection pressure and under a sufficient level of phenotypic noise (temperature). As the fraction of target spins is increased, the robust fitted phase is slightly expanded to a higher temperature, whereas higher selection pressure is needed to achieve it. The existence of an optimal fraction for attaining high fitness under intermediate phenotypic noise is suggested. This may explain why, in biological systems, such as in proteins, the fraction of units that are responsible for function is generally limited, and a sufficient fraction of nonfunctional units is needed, providing redundancy. While it is hard to obtain an accurate estimation of the functional sites in real proteins, rough estimates suggest they are of the order of 10–20% [47]. It hence will be interesting to discuss this range of functional regions, in relationship with evolvability and robustness as discussed here.

In the present theory, the proportionality between the standard thermodynamic susceptibility and mutational susceptibility is derived in the robust fitted phase. As the susceptibility measures the change of fitness due to varying conditions, a correlation between responses to environmental perturbations and that by genetic changes is suggested. Such correlation, or evolutionary fluctuation-response relationship [10,14,36–39] has been observed in experimental data from the evolution of protein dynamics and bacterial protein expressions, whereas we can derive it here under the replica symmetry assumption. As argued in [14], such correlation can only be achieved in the replica symmetric region where the original high-dimensional dynamics of the phenotypes are reduced to a low-dimensional manifold due to evolution towards robustness. The variations of fitness due to noise and that due to mutation then happen to occur along the same low-dimensional manifold, resulting in a correlation between them. If RSB occurs, such restriction of the phenotypic dynamics no longer exits, because in this case changes of fitness upon varying the environmental conditions will vary arbitrarily from realization to realization of the J_{ij}'s dynamics. As a result, the system will have random, uncorrelated responses to noise and to mutations.

In our formulation by assuming that the entire system reaches an equilibrium, we approximate the effect of the slowly evolving $\mathbf{J}^{(to)}$ that couple the subsystem $S_{\mathcal{T}}$ to $S_{\mathcal{O}}$ on these subsystems' own dynamics by the equilibrium correlations $\langle J^{(to)} J^{(to)} \rangle$. Such correlations are then incorporated separately into each of the dynamics of the $\mathbf{J}^{(tt)}$ and $\mathbf{J}^{(oo)}$ couplings by modifying the effective potentials of these dynamics, thus making the dynamics of these two different sets of coupling *independent* of each other. In an equilibrium statistical physics formulation, this leads to the necessity of introducing another type of replica, so-called coupling replicas, into the partition functions, besides the first (standard) spin replicas that take care of the effect of the phenotypes on the evolution of genotypes. This scheme hence allows us to treat the model in a standard mean-field manner. On one hand, being of mean-field nature, our approach can not provide a formal argument to support the hypothesis of [14] about the emergent dimensional reduction from phenotype-genotype coevolution. On the other hand, the correlation between the mutational and environmental susceptibilities in the robust fitted phase suggests the existence of a funnel-like landscape [48,49] that reinforces the dynamics to reside in a low-dimensional manifold by a global attraction.

Despite of its abstraction, our model may have some other implications on the proteins that acquire function through evolution. First, the presented correlation between robustness to noise and to mutation has been discussed previously in RNA [9] and in proteins [39]. Further quantitative analysis will be needed to establish the mechanism of such a general phenomenon. Second, the observed consistency between a global attraction to the robust fitted state in the replica symmetric phase and the funnel picture in protein folding dynamics suggests the importance of replica analysis in studying the protein free-energy landscape.

The current choice of fitness for the sake of simplicity, however, limits the possibility of having different global maxima in the fitness landscape. One can enrich the model behavior by determining fitness either by a combination of N_{fit} different target spin configurations or by a set of gauge-equivalent configurations.

In the present framework, since the couplings are symmetric, we constructed the effective potential of the coupling dynamics based partly on the existence of an energy landscape. For those models in other contexts [50,51] having such a landscape picture, we expect a straightforward application of our approach. Furthermore, the present double-replica theory can be extended to those stochastic dynamical systems that are not governed by Hamiltonian dynamics as well. In this case, instead of the effective potential and its associated partition function, one would need to characterize the ensemble of trajectories in the combined space of phenotypes and genotypes, using the moment generating function [52,53]. While we so far have solely used Hamiltonian dynamics as the main example of our approach, such an extension would allow for the applications to coevolution of gene-expression patterns and the gene-regulatory networks [10], that of species abundances and their ecological networks [54], and that of neuronal activities and network shaped by neural dynamics (learning) [55,56].

ACKNOWLEDGMENTS

We acknowledge support from Novo Nordisk Foundation (0065542) and would like to thank A. Sakata, K. Hukushima, Y. Kabashima, and Q.-Y. Tang for stimulating discussions.

APPENDIX A: DETAILS OF THE SHK MODEL

In the following we call the model originally introduced in [15,16] as the SHK model. In this model, phenotypes are spin configurations, and genotypes are the interaction matrix for spins. In a system of N spins, each spin i can take values $s_i \in \{-1, 1\}$ and is linked to exactly $N-1$ other spins,

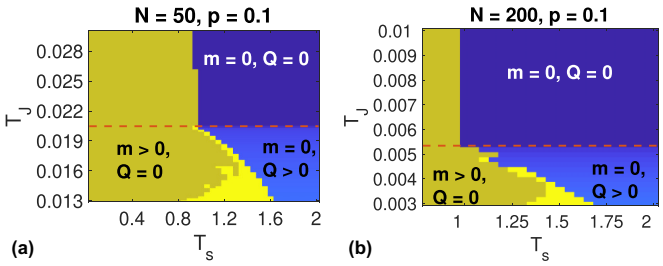

FIG. 5. The phase diagrams obtained by combining the behavior of m as function of T_s and T_J with that of Q at $p = 0.1$ for $N = 50$ (a) and $N = 200$ (b). The yellow region in both (a) and (b) shows the robust fitted phase ($m > 0$ and $Q \sim 1$). The red dashed line depicts the critical line $\beta_J^{(c)} = 1/(N - N_t)$ that is obtained from Landau's order-parameter-theory analysis of the free energy of target spins using the expression in Eq. (B4a).

thus forming a fully-connected network. Moreover, fitness is determined by a subset of target spins denoted by $\mathcal{T}$. Those spins that do not contribute to the fitness are called nontarget. The fitness Ψ at a noise level T_s is determined by the spin configurations at equilibrium as

$$\Psi(\mathbf{s}) = \frac{1}{N_t}\left\langle \left| \sum_{i\in\mathcal{T}} s_i \right| \right\rangle_{T_s}, \tag{A1}$$

where $\langle\cdot\rangle_{T_s}$ is the thermal average according to an equilibrium distribution over spin configurations only. This distribution is computed from the partition function of a spin-glass Hamiltonian $H_S = -\sum_{i<j} J_{ij} s_i s_j$ [24] in which the couplings J_{ij} are regarded as *fixed* over the course of the spin dynamics because they are assumed to evolve on much slower timescale than that of the spins. Here the couplings are symmetric, i.e., $J_{ij} = J_{ji}$, and are independently and identically distributed by a Gaussian distribution with zero mean and the variance $J^2 := \mathrm{var}(J_{ij}) = N^{-1}$. The model Hamiltonian of the full system is given by

$$H_S = -\sum_{i<j} J_{ij} s_i s_j. \tag{A2}$$

Once the spins have relaxed to an equilibrium at a temperature T_s via a Glauber update specified by H_S, the couplings are next updated with probability $\Pr[\mathbf{J} \to \tilde{\mathbf{J}}] = \min\{1, e^{\beta_J \Delta\Psi}\}$, where $\Delta\Psi = \Psi(\tilde{\mathbf{J}}) - \Psi(\mathbf{J})$ and $\beta_J \equiv 1/T_J$ is the genotypic selection pressure. These two dynamics are implemented consecutively one after another until the entire system equilibrates. Implementing this way, the model captures the evolution of feedback process between the phenotype and genotype, where the phenotype dynamics are represented by the stochastic dynamics of spins ($\mathbf{s}$) according to the energy landscape H_S for given genotype ($\mathbf{J}$), whereas the evolution of genotype is given by the stochastic change of ($\mathbf{J}$) according to the fitness $\Psi(\mathbf{s})$ determined by the phenotype. In contrast to more common theories of evolution, this model hence explicitly considers the co-evolution of these coupled landscapes.

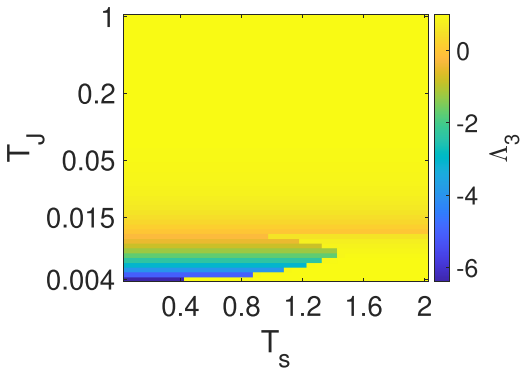

FIG. 7. The third-largest eigenvalue Λ_3 of the Hessian matrix for the *target* coupling replicas as function of the model parameter. Here we find the region with broken replica symmetry where $\Lambda_3 < 0$. The change in the sign of Λ_3 happens to coincide with the transition between $Q = 0$ and $Q > 0$. Here $N_t = 10$ and $N = 100$. Note the y axis is on logarithmic scale.

APPENDIX B: REPLICA SYMMETRIC ANSATZ SOLUTION AND THE EXPRESSION OF I_{kz} AND $\tilde{I}_k$

The partition functions are given in terms of the *target* and the *nontarget* free energy densities, $f_{\mathcal{T}}(\mathbf{m}, \mathbf{q}, \mathbf{r}, \mathbf{Q}, \mathbf{M})$ and $f_{\mathcal{O}}(\tilde{\mathbf{m}}, \tilde{\mathbf{q}}, \tilde{\mathbf{r}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{M}})$, respectively, by

$$\mathcal{Z}_{\mathcal{T}} = \int D\mathbf{m}\, D\mathbf{q}\, D\mathbf{r}\, D\mathbf{Q}\, D\mathbf{M}\, e^{-\beta_J N p f_{\mathcal{T}}(\mathbf{m},\mathbf{q},\mathbf{r},\mathbf{Q},\mathbf{M})}, \tag{B1a}$$

$$\mathcal{Z}_{\mathcal{O}} = \int D\tilde{\mathbf{m}}\, D\tilde{\mathbf{q}}\, D\tilde{\mathbf{r}}\, D\tilde{\mathbf{Q}}\, D\tilde{\mathbf{M}}\, e^{-\beta_J N(1-p) f_{\mathcal{O}}(\tilde{\mathbf{m}},\tilde{\mathbf{q}},\tilde{\mathbf{r}},\tilde{\mathbf{Q}},\tilde{\mathbf{M}})}, \tag{B1b}$$

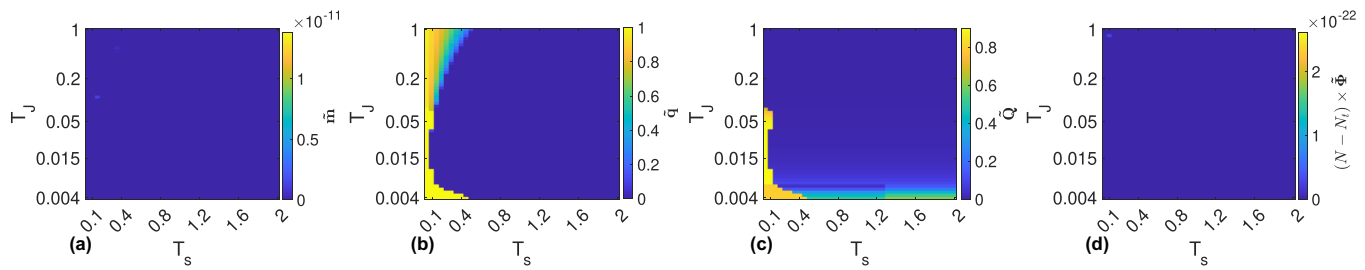

FIG. 6. Magnetization for nontarget spins $\tilde{m}$ (a). Overlap between different replicas for nontarget spins $\tilde{q}$ (b). Averaged correlation of a pair of couplings between a target and a nontarget spin that share a common target spin $\tilde{Q}$ (c). Averaged value of the link $J^{(oo)}$ among nontarget spins $\tilde{\Phi}$ (d). Here $N_t = 10$, $N = 100$. Note the y axis is on logarithmic scale.

where

$$f_{\mathcal{T}} = \frac{1}{2}\left\{\sum_{a<b}\frac{q_{ab}^2}{n^2} + \sum_{k<k'}\frac{Q_{k,k'}^2}{(N-t)^2} + \frac{1}{n(N-t)}\sum_{a,k}M_{ak}^2\right\} - \frac{1}{\beta_J}\ln\sum_{\{s_i;s^\alpha;\sigma^k\}_{i\in\mathcal{T}}} e^L, \tag{B2a}$$

$$f_{\mathcal{O}} = \frac{1}{2}\left\{\tilde{K}\sum_a\frac{\tilde{m}_a^2}{n} + \tilde{K}\sum_k\frac{\tilde{r}_k^2}{t} + \sum_{a<b}\frac{\tilde{q}_{ab}^2}{n^2} + \sum_{k<k'}\frac{\tilde{Q}_{k,k'}^2}{t^2} + \frac{1}{nt}\sum_{a,k}\tilde{M}_{ak}^2\right\} - \frac{1}{\beta_J}\ln\sum_{\{s^\alpha;\sigma^k\}} e^{\tilde{L}}, \tag{B2b}$$

$$L = \frac{\beta_J}{4N_t}\left|\sum_{i\in\mathcal{T}}s_i\right|^2 - \frac{\beta_J}{2}\left|\sum_{i\in\mathcal{T}}s_i\right|\left[\sum_{a=1}^{n}\frac{m_a^2}{n} + \sum_{k}^{N-N_t}\frac{r_k^2}{N-N_t}\right]$$
$$+\beta_J\left(\frac{1}{N_t}\left|\sum_{i\in\mathcal{T}}s_i\right|\left[\sum_{a=1}^{n}\frac{m_a s^a}{n} + \sum_{k}^{N-N_t}\frac{r_k\sigma^k}{N-N_t}\right] + \sum_{a<b}^{n}\frac{q_{ab}s^a s^b}{n^2} + \sum_{k<k'}^{N-N_t}\frac{Q_{kk'}\sigma^k\sigma^{k'}}{(N-N_t)^2} + \sum_{a,k}\frac{M_{ak}s^a\sigma^k}{n(N-N_t)}\right), \tag{B3a}$$

$$\tilde{L} = \beta_J\left(\frac{\tilde{K}}{n}\sum_{a=1}^{n}\tilde{m}_a s^a + \frac{\tilde{K}}{N_t}\sum_{k}^{N_t}\tilde{r}_k\sigma^k + \frac{1}{n^2}\sum_{a<b}^{n}\tilde{q}_{ab}s^a s^b + \frac{1}{N_t^2}\sum_{k<k'}^{N_t}\tilde{Q}_{kk'}\sigma^k\sigma^{k'} + \frac{1}{nN_t}\sum_{a,k}\tilde{M}_{ak}s^a\sigma^k\right). \tag{B3b}$$

Denoting $\mathcal{D}x\,\mathcal{D}y = \frac{e^{-(x^2+y^2)/2}}{2\pi}\,dx\,dy$ and $A_z(m,r) = \frac{\beta_J}{4}\frac{(N_t-2z)^2}{N_t^2} - \frac{\beta_J}{2}\frac{|N_t-2z|}{N_t}(m^2+r^2)$, we have

$$I_{k,z} = \int \mathcal{D}x\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\frac{|1-2z|}{N_t}\right)\right\}\left[\cosh\left(\beta_s m\frac{|N_t-2z|}{N_t} + x\frac{\sqrt{\beta_J q}}{n} + \frac{\beta_J M}{n}\frac{N-N_t-2k}{N-N_t}\right)\right]^n \tag{B4a}$$

$$\tilde{I}_k = \int \mathcal{D}x\mathcal{D}y\,\exp\left\{\frac{N_t-2k}{N_t}\left(y\sqrt{\beta_J\tilde{Q}} + \beta_J\tilde{K}\tilde{r}\right)\right\}\left[\cosh\left(\beta_s\tilde{K}\tilde{m} + x\frac{\sqrt{\beta_J\tilde{q}}}{n} + \frac{\beta_J\tilde{M}}{n}\frac{N_t-2k}{N_t}\right)\right]^n. \tag{B4b}$$

The argument of the $\cosh(\cdot)$ function will be denoted by

$$\Omega = \beta_s m\frac{|N_t-2z|}{N_t} + x\frac{\sqrt{\beta_J q}}{n} + \frac{\beta_J M}{n}\frac{N-N_t-2k}{N-N_t}, \quad \tilde{\Omega} = \beta_s\tilde{K}\tilde{m} + x\frac{\sqrt{\beta_J\tilde{q}}}{n} + \frac{\beta_J\tilde{M}}{n}\frac{N_t-2k}{N_t}. \tag{B5}$$

The replica symmetric free energy densities given in the main text yield the extremum condition after setting $\tilde{K}=0$:

$$m = \frac{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega)]^n\tanh(\Omega)}{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega(x,q,M))]^n},$$

$$q = \frac{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega)]^n[\tanh(\Omega)]^2}{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega)]^n},$$

$$Q = -\frac{1}{N-N_t} + \frac{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\left(\frac{N-N_t-2k}{N-N_t}\right)^2 I_{k,z}}{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}I_{k,z}}, \quad r = \frac{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\frac{N-N_t-2k}{N-N_t}I_{k,z}}{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}I_{k,z}},$$

$$M = \frac{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\frac{N-N_t-2k}{N-N_t}\int \mathcal{D}x\,\mathcal{D}y\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega]^n\tanh(\Omega)}{\sum_{z=0}^{N_t}\binom{N_t}{z}\sum_{k=0}^{N-N_t}\binom{N-N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{A_z + \frac{N-N_t-2k}{N-N_t}\left(y\sqrt{\beta_J Q} + \beta_J r\theta_z\right)\right\}[\cosh(\Omega)]^n},$$

$$\tilde{m} = \frac{\sum_{k=0}^{N_t}\binom{N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n\tanh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))}{\sum_{k=0}^{N_t}\binom{N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n},$$

$$\tilde{q} = \frac{\sum_{k=0}^{N_t}\binom{N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n[\tanh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^2}{\sum_{k=0}^{N_t}\binom{N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n},$$

$$\tilde{Q} = -\frac{1}{N_t} + \frac{\sum_{k=0}^{N_t}\binom{N_t}{k}\left(\frac{N_t-2k}{N_t}\right)^2\tilde{I}_k}{\sum_{k=0}^{N_t}\binom{N_t}{k}\tilde{I}_k}, \quad \tilde{r} = \frac{\sum_{k=0}^{N_t}\binom{N_t}{k}\frac{N_t-2k}{N_t}\tilde{I}_k}{\sum_{k=0}^{N_t}\binom{N_t}{k}\tilde{I}_k},$$

$$\tilde{M} = \frac{\sum_{k=0}^{N_t}\binom{N_t}{k}\frac{N_t-2k}{N_t}\int \mathcal{D}x\,\mathcal{D}y\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n\tanh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))}{\sum_{k=0}^{N_t}\binom{N_t}{k}\int \mathcal{D}x\,\mathcal{D}y\,\exp\left\{\frac{N_t-2k}{N_t}y\sqrt{\beta_J Q}\right\}[\cosh(\Omega(x,\tilde{m},\tilde{q},\tilde{M}))]^n}.$$

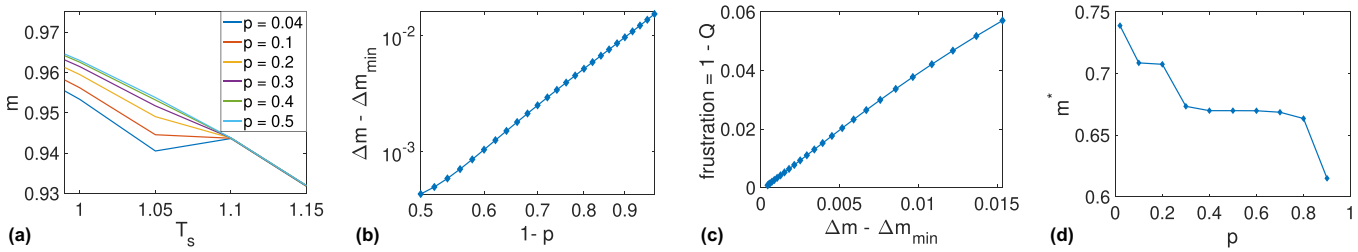

FIG. 8. (a) A zoom-in of Fig. 4(a) in the region nearby $T_c^{(1)}$. (b) The drop of fitness subtracted from the minimal value $\Delta m_{\min}$ as a function of the fraction of nontarget spins. Here $\Delta m_{\min}$ is defined as the drop of magnetization at the critical number of target spins $N_t^{(c)} = 60$ at $T_J = 0.005$, while $\Delta m = 1 - m(T_c^{(1)})$ is the drop of of fitness at $T_c^{(1)}$. (c) Linear relationship between the frustration $1 - Q$ and $\Delta m - \Delta m_{\min}$ at $T_c^{(1)}$. (d) The drop m^* of fitness at $T_c^{(2)}$. Here $N = 100$.

APPENDIX C: N DEPENDENCE OF THE PHASE DIAGRAM

Using an expansion of the free energy density of the target spins, based on Eq. (B4a), we can estimate the highest value of T_J above which Q drops from a high value to zero: $\beta_J^{(c)} = 1/(N - N_t)$. Therefore, the robust fitted phase necessarily exists for all values of N, though, at a given fraction $p = N_t/N$, a higher selection pressure, i.e., a lower T_J is needed to achieve it for larger N. We show this kind of N dependence of the phase diagram for $N = 50$ and $N = 200$ in Fig. 5. Overall, the structure of the phase diagram remains the same as that obtained for $N = 100$ in Fig. 3(a) of the main text. The difference in the exact location of the highest value of T_J above which Q drops from a high value to zero is well explained by our estimation shown as a dashed horizontal line.

APPENDIX D: PHASE DIAGRAM OF THE ORDER PARAMETERS OF THE NONTARGET SPINS AND THAT OF THE EIGENVALUE OF THE HESSIAN

We present the phase diagram for the order parameters of the non-target spins in Fig. 6 and that for the third-largest eigenvalue of the Hessian Λ_3 in Fig. 7. Without any effect of fitness acting on them, nontarget spins have only zero magnetization [Fig. 6(a)] and the averaged value $\tilde{\Phi}$ of the coupling among them $J^{(oo)}$ is equal to zero [Fig. 6(d)]. Nevertheless, due to the effect of random couplings, $J^{(oo)}$, a spin-glass ordering can appear at low T_J and T_s as seen in Fig. 6(c). As we expect a dependence of the drop in fitness at $T_c^{(1)}$, which is denoted by $\Delta m = 1 - m$, on the fraction of target spins, we first show a zoom-in of the behavior of m at T_s close to $T_c^{(1)}$ in Fig. 8(a). We then find that at the critical number of target spins $N_t^{(c)}$, this change has a minimal value $\Delta m_{\min}$, so that $\Delta_m - \Delta m_{\min}$ can be depicted as function of $1 - p$ in Fig. 8(b). Once subtracted Δm from $\Delta m_{\min}$, we find a linear relationship exists between $\Delta m - \Delta m_{\min}$ and $1 - Q$. Such relationship is demonstrated Fig. 8(c). We finally measure how much fitness changes under a transition from robust to paramagnet phase at $T_c^{(2)}$. We denote this kind of drop by m^* in Fig. 8(d), where we find a decrease of m^* with increasing p, implying that fitness is more robust at higher p.

[1] S. Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution, *Proceedings of the XI International Congress of Genetics*, Vol. 1 (1932), pp. 356–366.

[2] J. A. G. de Visser and J. Krug, Empirical fitness landscapes and the predictability of evolution, Nat. Rev. Genet. **15**, 480 (2014).

[3] D. Nichol, M. Robertson-Tessi, A. R. A. Anderson, and P. Jeavons, Model genotype–phenotype mappings and the algorithmic structure of evolution, J. R. Soc. Interface **16**, 20190332 (2019).

[4] H. H. McAdams and A. Arkin, Stochastic mechanisms in gene expression, Proc. Natl. Acad. Sci. USA **94**, 814 (1997).

[5] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Stochastic gene expression in a single cell, Science **297**, 1183 (2002).

[6] C. Furusawa, T. Suzuki, A. Kashiwagi, T. Yomo, and K. Kaneko, Ubiquity of log-normal distributions in intra-cellular reaction dynamics, Biophysics **1**, 25 (2005).

[7] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai, Noise in protein expression scales with natural protein abundance, Nat. Genet. **38**, 636 (2006).

[8] A. Wagner, *Robustness and Evolvability in Living Systems*, Princeton Studies in Complexity Vol. 24 (Princeton University Press, 2013).

[9] L. W. Ancel and W. Fontana, Plasticity, evolvability, and modularity in RNA, J. Exp. Zool. **288**, 242 (2000).

[10] K. Kaneko, Evolution of robustness to noise and mutation in gene expression dynamics, PLoS ONE **2**, e434 (2007).

[11] Z. Shreif and V. Periwal, A network characteristic that correlates environmental and genetic robustness, PLoS Comput. Biol. **10**, e1003474 (2014).

[12] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics*, Springer Series in Solid-State Sciences (Springer, Berlin, 1985).

[13] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, Oxford Master Series in Physics Vol. 14 (Oxford University Press, 2021).

[14] A. Sakata and K. Kaneko, Dimensional Reduction in Evolving Spin-Glass Model: Correlation of Phenotypic Responses to Environmental and Mutational Changes, Phys. Rev. Lett. **124**, 218101 (2020).

[15] A. Sakata, K. Hukushima, and K. Kaneko, Funnel Landscape and Mutational Robustness as a Result of Evolution under Thermal Noise, Phys. Rev. Lett. **102**, 148101 (2009).

[16] A. Sakata, K. Hukushima, and K. Kaneko, Statistical-mechanical study of evolution of robustness in noisy environments, Phys. Rev. E **80**, 051919 (2009).

[17] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1986).

[18] R. W. Penney, A. C. C. Coolen, and D. Sherrington, Coupled dynamics of fast spins and slow interactions in neural networks and spin systems, J. Phys. A: Math. Gen. **26**, 3681 (1993).

[19] A. C. C. Coolen, R. W. Penney, and D. Sherrington, Coupled dynamics of fast spins and slow interactions: An alternative perspective on replicas, Phys. Rev. B **48**, 16116 (1993).

[20] R. Penney and D. Sherrington, Slow interaction dynamics in spin-glass models, J. Phys. A: Math. Gen. **27**, 4027 (1994).

[21] V. Dotsenko, S. Franz, and M. Mezard, Partial annealing and overfrustration in disordered systems, J. Phys. A: Math. Gen. **27**, 2351 (1994).

[22] T. Uezu, K. Abe, S. Miyoshi, and M. Okada, Statistical mechanical study of partial annealing of a neural network model, J. Phys. A: Math. Theor. **43**, 025004 (2010).

[23] A. Sakata, K. Hukushima, and K. Kaneko, Replica symmetry breaking in an adiabatic spin-glass model of adaptive evolution, Europhys. Lett. **99**, 68004 (2012).

[24] D. Sherrington and S. Kirkpatrick, Solvable Model of a Spin-Glass, Phys. Rev. Lett. **35**, 1792 (1975).

[25] Note the difference from the model implementation of [15] where couplings are updated at discrete time steps.

[26] Technically speaking, the field K is introduced to break the gauge-symmetry between different local Mattis states that have no frustration among target spins (i.e., triples $J_{ij}^{(tt)}J_{jk}^{(tt)}J_{ki}^{(tt)} > 0$, $\forall\, i < j < k \in \mathcal{T}$). As this symmetry is broken, the ferromagnetic state (with all links between target spins being positive) is distinguished from all other unfrustrated states and, as a consequence, alignment among target spins can be achieved at low T_s. One might want to consider a different form of K that explicitly takes into account the Boltzmann weights corresponding to the Hamiltonian H_S [i.e., $e^{-\beta_s H_S}$, as defined in Eq. (4)]. Such a term, however, may not guarantee the evolution towards a subgraph of only ferromagnetic interactions among target spins.

[27] Here we used the identities $\langle s_i s_j\rangle_{T_s} = \frac{1}{\beta_s}\frac{\partial}{\partial J_{ij}^{(tt)}}\ln Z_1$ and $\langle s_k s_{k'}\rangle_{T_s} = \frac{1}{\beta_s}\frac{\partial}{\partial J_{kk'}^{(oo)}}\ln \tilde{Z}_1$, where the partition function of the target spins' subsystem and that of the nontarget spins' subsystem are $Z_1 := \sum_{\{s_i\}_{i\in\mathcal{T}}} e^{-\beta_s H_{\mathcal{T}}(\mathbf{J}^{(tt)})}$ and $\tilde{Z}_1 := \sum_{\{s_i\}_{i\notin\mathcal{T}}} e^{-\beta_s H_{\mathcal{O}}(\mathbf{J}^{(oo)})}$, respectively.

[28] Requiring that the couplings should remain bounded as $\tau \to \infty$, we need to introduce a decay term $-\lambda_{ij}J_{ij}$ to the couplings dynamics. In order to keep the genotypic selection pressure the same for all the couplings regardless of their types, we choose $-\lambda_{ij}J_{ij}^{(tt)} = -\sqrt{p}J_{ij}^{(tt)}$ and $-\lambda_{ij}J_{ij}^{(oo)} = -\sqrt{1-p}J_{ij}^{(oo)}$, where $p = N_t/N$ is the fraction of target spins.

[29] Note that $\mathcal{Z}_{\mathcal{T}}$ and $\mathcal{Z}_{\mathcal{O}}$ are different from the partition functions of the target spins' subsystem $Z_1(\mathbf{J}^{(tt)})$ and that of the nontarget spins' subsystem $\tilde{Z}_1(\mathbf{J}^{(oo)})$ given in [27].

[30] This can already be seen directly from Eq. (5) as the expression inside the curly braces does not depend on $J_{ij}^{(oo)}$ and $J_{ij}^{(to)}$.

[31] Though n appears as an integer number here, we will analytically continue to real positive n in subsequent steps of calculations.

[32] Here $I_{kz} = I_{kz}(m, q, Q, r, M)$ and $\tilde{I}_k = \tilde{I}_k(\tilde{m}, \tilde{q}, \tilde{Q}, \tilde{r}, \tilde{M})$ with $r, M, \tilde{r}, \tilde{M}$ are other variables that the free energy densities depend on. Since these observables for target and nontarget spins have their behavior correlated to that of (m, q, Q) and $(\tilde{m}, \tilde{q}, \tilde{Q})$, respectively, they do not provide additional information about the structure of the model phase diagram.

[33] J. R. L. de Almeida and D. J. Thouless, Stability of the sherrington-kirkpatrick solution of a spin glass model, J. Phys. A: Math. Gen. **11**, 983 (1978).

[34] The eigenvalue of the Hessian changes its sign on the border between these phases; see Appendix D. Note that though we do not calculate the full hierarchy of replica symmetry breaking (RSB) à la Parisi, but as long as RSB happens this nonrobustness is true for the full RSB solution.

[35] This optimality, however, does not happen in the spinglass phase **SP2** with low temperature T_s, since the selection pressure needed to achieve nonzero fitness crucially depends on p. In fact as p is increased, higher pressure (i.e., lower T_J) is needed to achieve nonzero m of targets, i.e., to transition from **SP1** to **SP2**.

[36] K. Sato, Y. Ito, T. Yomo, and K. Kaneko, On the relation between fluctuation and response in biological systems, Proc. Natl. Acad. Sci. USA **100**, 14086 (2003).

[37] K. Kaneko and C. Furusawa, An evolutionary relationship between genetic variation and phenotypic fluctuation, J. Theor. Biol. **240**, 78 (2006).

[38] S. Ciliberti, O. C. Martin, and A. Wagner, Robustness can evolve gradually in complex regulatory gene networks with varying topology, PLoS Comput. Biol. **3**, e15 (2007).

[39] Q.-Y. Tang and K. Kaneko, Dynamics-Evolution Correspondence in Protein Structures, Phys. Rev. Lett. **127**, 098103 (2021).

[40] C. R. Landry, B. Lemos, S. A. Rifkin, W. J. Dickinson, and D. L. Hartl, Genetic properties influencing the evolvability of gene expression, Science **317**, 118 (2007).

[41] M.-A. Félix and M. Barkoulas, Pervasive robustness in biological systems, Nat. Rev. Genet. **16**, 483 (2015).

[42] R. Silva-Rocha and V. de Lorenzo, Noise and robustness in prokaryotic regulatory networks, Annu. Rev. Microbiol. **64**, 257 (2010).

[43] Y. Uchida, S. Shigenobu, H. Takeda, C. Furusawa, and N. Irie, Potential contribution of intrinsic developmental stability toward body plan conservation, BMC Biology **20**, 82 (2022).

[44] More precisely, in [14] $\Psi_i = \langle \mathrm{sign}(m) s_i\rangle_{T_s}$.

[45] From the replica-symmetric free energy density $f_{\mathcal{T}}^{\mathrm{RS}}$ we get $\chi_m^{\mathrm{RS}} = \beta_s[1 + (n-1)q]$, while M can not be obtained directly within this ansatz.

[46] B. Lehner, Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast, PLoS ONE **5**, 1 (2010).

[47] N. Wu, S. N. Yaliraki, and M. Barahona, Prediction of protein allosteric signalling pathways and functional residues through paths of optimised propensity, J. Mol. Biol. **434**, 167749 (2022).

[48] N. Go, Theoretical studies of protein folding, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).

[49] J. N. Onuchic and P. G. Wolynes, Theory of protein folding, Curr. Opin. Struct. Biol. **14**, 70 (2004).

[50] F. C. Poderoso and J. F. Fontanari, Model ecosystem with variable interspecies interactions, J. Phys. A: Math. Theor. **40**, 8723 (2007).

[51] M. T. Pham, I. Kondor, R. Hanel, and S. Thurner, The effect of social balance on social fragmentation, J. R. Soc. Interface **17**, 20200752 (2020).

[52] P. C. Martin, E. D. Siggia, and H. A. Rose, Statistical dynamics of classical systems, Phys. Rev. A **8**, 423 (1973).

[53] C. De Dominicis and L. Peliti, Field-theory renormalization and critical dynamics above T_c: Helium, antiferromagnets, and liquid-gas systems, Phys. Rev. B **18**, 353 (1978).

[54] M. Barbier, J.-F. Arnoldi, G. Bunin, and M. Loreau, Generic assembly patterns in complex ecological communities, Proc. Natl. Acad. Sci. USA **115**, 2156 (2018).

[55] J. Kadmon and H. Sompolinsky, Transition to Chaos in Random Neuronal Networks, Phys. Rev. X **5**, 041030 (2015).

[56] J. Schuecker, S. Goedeke, and M. Helias, Optimal Sequence Memory in Driven Random Networks, Phys. Rev. X **8**, 041029 (2018).