Journal of Physics A: Mathematical and Theoretical

PAPER • OPEN ACCESS

Ultrametric identities in glassy models of natural evolution

To cite this article: Elena Agliari *et al* 2023 *J. Phys. A: Math. Theor.* **56** 385001

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **56** (2023) 385001 (24pp) https://doi.org/10.1088/1751-8121/acf101

Ultrametric identities in glassy models of natural evolution

Elena Agliari[1,2,*], **Francesco Alemanno**[3,4], **Miriam Aquaro**[1,2] and **Adriano Barra**[2,3]

[1] Dipartimento di Matematica, Sapienza Università di Roma, Roma, Italy
[2] GNFM-INdAM, Gruppo Nazionale di Fisica Matematica, Istituto Nazionale di Alta Matematica, Italy
[3] Dipartimento di Matematica e Fisica, Università del Salento, Lecce, Italy
[4] Dipartimento di Matematica, Università di Bologna, Bologna, Italy

E-mail: agliari@mat.uniroma1.it

Received 21 March 2023; revised 18 July 2023
Accepted for publication 16 August 2023
Published 29 August 2023

Abstract

Spin-glasses constitute a well-grounded framework for evolutionary models. Of particular interest for (some of) these models is the lack of self-averaging of their order parameters (e.g. the Hamming distance between the genomes of two individuals), even in asymptotic limits, much as like what happens to the overlap between the configurations of two replica in mean-field spin-glasses. In the latter, this lack of self-averaging is related to a peculiar behavior of the overlap fluctuations, as described by the Ghirlanda–Guerra identities and by the Aizenman–Contucci polynomials, that cover a pivotal role in describing the ultrametric structure of the spin-glass landscape. As for evolutionary models, such identities may therefore be related to a taxonomic classification of individuals, yet a full investigation on their validity is missing. In this paper, we study ultrametric identities in simple cases where solely random mutations take place, while selective pressure is absent, namely in *flat landscape* models. In particular, we study three paradigmatic models in this setting: the *one parent model* (which, by construction, is ultrametric at the level of single individuals), the *homogeneous population model* (which is replica symmetric), and the *species formation model* (where a broken-replica scenario emerges at the level of species). We find analytical and numerical evidence that in the first and in the third model nor the Ghirlanda–Guerra neither the Aizenman–Contucci

[*] Author to whom any correspondence should be addressed.

constraints hold, rather a new class of ultrametric identities is satisfied; in the second model all these constraints hold trivially. Very preliminary results on a real biological human genome derived by *The 1000 Genome Project Consortium* and on two artificial human genomes (generated by two different types neural networks) seem in better agreement with these new identities rather than the classic ones.

Keywords: glassy evolutionary models, ultrametric identities,
flat and rugged landscapes

(Some figures may appear in colour only in the online journal)

1. Introduction

1.1. The evolutionary interpretation of spin glasses

This special issue of the Journal of Physics A, dedicated to *Random landscapes and dynamics in evolution, ecology and beyond* to celebrate the Nobel Prize awarded to Giorgio Parisi in 2021, highlights the widespread applications of spin glasses. Indeed, spin glasses constitute a paradigmatic example of complex system [1–3] and their peculiar behavior is often evoked when describing some non-trivial phenomenologies occurring in disparate areas of Science, from several branches of Biology (e.g. neurology [4–6], genomics [7, 8], immunology [9–11], ecology [12–14]) to Sociology [15, 16], Economics [17, 18], Computer Science [19, 20] and more. In particular, quoting a sentence of Parisi's Nobel Lecture 'Multiple Equilibria': *Ultrametricity and taxonomy are essentially related: a standard taxonomy, i.e. a hierarchical classification, is possible only if the relevant properties have an ultrametric structure. In the standard taxonomic classification of living beings, the distance is related to the history of evolution. In spin glasses, the taxonomy is intrinsic to the static equilibrium properties of the system and it is not related to evolution in time.*

In this paper, we just focus on the framework provided by spin-glass theory to Natural Evolution, which has attracted much interest in the past decades and nowadays represents an insightful and solid branch of modern disordered statistical mechanics; specifically, we will deepen the relation between taxonomic classification and spin glasses.

In the picture of Evolutionary Biology, genomic randomness plays a crucial role; take for instance the *adaptive walks* approach, where Natural Evolution is modeled as a two-step stochastic process: *i*. the genotype of a species undergoes random mutations, *ii*. its newborns with higher fitness are preserved (see e.g. [21]). Then, as pointed out by Eigen, the compromise between replication efficiency and frequency of mutations in evolutionary dynamics is conceptually close to the compromise between energy minimization and entropy maximization in statistical mechanics, moreover, the *error threshold* in the mutation rate in the former mimics the thermal noise of the latter [22, 23]. Also, to quantify the genetic variability within a population, we can introduce a proximity measure q between pairs of individuals that plays like an order parameter (mirroring the replica overlap) and, just like in disordered systems, two distinct averages can be implemented, namely the *population average* (mirroring the thermal average) and the *process average* (mirroring the quenched average) [24]. Therefore, one can consider the population-average of q, which in general depends on time, and inspect whether its average over a long time stretch exhibits vanishing fluctuations. If this is the case we have a *self-averaging* structure of the model (in such a way that its main features could be captured by the quasi-species[5] limiting description), and, if not, we have a *non-self-averaging* structure, that is a hallmark of complex systems.

In this context it is also worth recalling Wright's Adaptive Landscape (see e.g. [30]), Fisher's Fundamental Theorem of Natural Selection, Kimura's Neutral Theory [31], and Felsenstein's statistical methods to reconstruct evolutionary trees [32, 33], which constitute fundamental steps [34]. Along these lines, the development of a disordered statistical mechanical theory for Natural Evolution was started by Leuthäusser [35] and Tarazona [23] and a spin-glass setting was pursued by Derrida, Higgs, Franz, Peliti, Sellitto, and Serva, just to name a few (see e.g. [22, 24, 36–39] and references therein). More specifically, we recognize two classes of models: those where both random mutations and selective pressure are involved (also referred to as *rugged landscape* models) and those where evolution is driven only by random mutations (also referred to as *flat landscape* models). Reference models for the former are the P-spin-glass [21], the random energy model (REM) [22] and the Hopfield model [23], while for the latter we mention the one parent model (OPM) [24], the homogeneous population model (HPM) [38], and the species formation model (SFM) [36, 37]. The OPM is asexuated and its order parameter lacks self-averaging, while its sexuated counterpart, the HPM, is self-averaging, unless a threshold in the similarity between the two genomes that are matching to reproduce is introduced and this case corresponds to the SFM. Notably, in the latter, the presence of a similarity threshold yields a persistent, spontaneous formation and extinction process at the level of species with consequent breakdown of self-averaging.

The behavior of the order-parameter fluctuations in spin-glass models has been extensively studied, starting from the fully-connected Sherrington–Kirkpatrick (SK) model [40–42], to its generalizations (see e.g. [43–60] and section 1.2 for more details), including the rugged landscape models mentioned above (where both random mutations and selective pressure are at work). There, the order parameters are proved to be non-self-averaging and their momenta satisfy a class of non-trivial identities known as Ghirlanda–Guerra and Aizenman–Contucci (the latter are actually a family of identities that is a subset of the Ghirlanda–Guerra ones). We recall that Ghirlanda–Guerra identities played a pivotal role in Panchenko's proof of Parisi ultrametricity in the SK model [61–63] and ultrametricity, in turns, covers a key role in Natural Evolution (think for instance at the taxonomic classification in Biology).

In the context of Natural Evolution, the presence of both random mutation and selective pressure seems thus to be associated to the break-down of self-averaging with the momenta of the order parameter obeying some constraints. However, the validity of such constraints in the case of flat landscape models is still an open question that deserves attention. In this work we prove analytically for the OPM the validity of a new class of identities and find numerical evidence for their validity also for the SFM for which, instead, classic identities seem to be violated.

Further, as a proof of concept, we test the standard constraints (both Ghirlanda–Guerra and Aizenman–Contucci) as well as our new identities, on a biological dataset (the *1000 human genome project* [64] as well as on synthetic datasets generated by two different neural networks (a Generative Adversarial Network, GAN [65], and a Restricted Boltzmann Machine, RBM [66]) obtaining in all these cases that the new identities are satisfied while the standard ones are mildly violated.

[5] Eigen's *quasi species* approach (see e.g. [25, 26]) neglects, by definition, fluctuations and evolution is ruled by deterministic equations reminiscent of reaction kinetics; see also [27–29] for a systematic formalization of reaction kinetics via statistical mechanics.

1.2. The harmonic oscillator of spin glasses: SK model

The SK model [1, 2, 67, 68] is defined in terms of the pairwise Hamiltonian

$$\mathcal{H}_N(\boldsymbol{S}|\boldsymbol{J}) = \frac{1}{\sqrt{N}}\sum_{i<j}^{N,N} J_{ij} S_i S_j, \tag{1.1}$$

where the symmetric couplings $\boldsymbol{J} = \{J_{ij}\}_{i<j}^{N,N}$ are $N(N-1)/2$ i.i.d. random variables sampled from $\mathbb{P}(J_{ij}) = \mathcal{N}(0,1)$[6] and the interacting units are N Ising spins $\boldsymbol{S} = \{S_1,\ldots,S_N\} \in \{-1,+1\}^N$.

For a given inverse temperature $\beta \in \mathbb{R}^+$ and for a quenched coupling setting $\boldsymbol{J}$, we introduce the Boltzmann-Gibbs measure $\mathcal{P}_{N,\beta}(\boldsymbol{S}|\boldsymbol{J})$, the partition function $\mathcal{Z}_{N,\beta}(\boldsymbol{J})$, and the quenched free-energy $\mathcal{F}_{N,\beta}$ that read as

$$\mathcal{P}_{N,\beta}(\boldsymbol{S}|\boldsymbol{J}) = \frac{\exp\left(-\beta\mathcal{H}_N(\boldsymbol{S}|\boldsymbol{J})\right)}{\mathcal{Z}_{N,\beta}(\boldsymbol{J})}, \tag{1.2}$$

$$\mathcal{Z}_{N,\beta}(\boldsymbol{J}) = \sum_{\{S\}}^{2^N} \exp\left(-\beta\mathcal{H}_N(\boldsymbol{S}|\boldsymbol{J})\right), \tag{1.3}$$

$$\mathcal{F}_{N,\beta} = \frac{1}{N}\mathbb{E}\ln\mathcal{Z}_{N,\beta}(\boldsymbol{J}), \tag{1.4}$$

where the expectation $\mathbb{E}$ is over the possibile realizations of $\boldsymbol{J}$ drawn from $\mathbb{P}$. Next, for a generic observable $O(\boldsymbol{S})$, we define the following averages

$$\omega_{N,\beta,\boldsymbol{J}}(O) := \sum_{\{S\}}^{2^N} O(\boldsymbol{S})\mathcal{P}_{N,\beta}(\boldsymbol{S}|\boldsymbol{J}) \tag{1.5}$$

$$\langle O\rangle_{N,\beta} := \mathbb{E}[\omega_{N,\beta,\boldsymbol{J}}(O)]. \tag{1.6}$$

Due to frustration among the spins in the network, once the temperature is lowered beyond a critical one $1/\beta_c$, the free-energy landscape of this system gets spontaneously rugged and minima hierarchically split one into another recursively; consequently, spins tend to freeze in configurations displaying no long-range ferromagnetic-like order. Then, a natural measure of (any) internal organization of the system is a similarity measure between the spin configurations obtained for two replicas of the system characterized by the same realization of disorder $\boldsymbol{J}$, namely two configurations sampled from the same distribution $\mathcal{P}_{N,\beta}(\boldsymbol{S}|\boldsymbol{J})$. In particular, the (simplest) order parameter is the two-replica overlap

$$q_{ab} := \frac{1}{N}\sum_i S_i^a S_i^b, \tag{1.7}$$

that is nothing but the normalized scalar product between the configurations, corresponding to two replicas labeled as a, b, and denoted as $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$. In the high-temperature region, spins behave independently of each other, replica configurations are uncorrelated and the overlap distribution is a Dirac delta peaked at zero, however, beyond β_c and in the thermodynamic limit $N \to \infty$, there emerge non-zero values for q_{ab}, such that the overlap distribution is a

[6] Beyond sampling from standard Gaussians, the couplings can be drawn with Rademacher entries and the same picture would be the same.

(possibly infinite) sum of Dirac's deltas at these values (the so-called Parisi plateau), and the whole distribution $\mathcal{P}_\beta(q)$ is retained as order parameter.

Thus, as ergodicity breaks down, this model breaks also the permutational invariance among its replicas giving rise to the well-known phenomenon of replica symmetry breaking (RSB): this was suggested as an ansatz by Giorgio Parisi in the eighties [1, 2] and then mathematically proved twenty years laters by Francesco Guerra [69] and Michel Talagrand [70] as for the expression of free energy and by Dmitry Panchenko [61–63] as for the hierarchical organization of its valleys (i.e. ultrametricity). Remarkably, Panchenko's proof is significantly based on the peculiar fluctuations of the overlap as summarized by the Ghirlanda-Guerra identities [42], *vide infra*. Indeed, among the most striking features of the emergent order of the SK model at low temperature lies the spontaneous ultrametric organization of its pure states, resembling taxonomic ordering in Natural Evolution, as for instance captured by the 3-replicas and 4-replicas overlap joint distributions $\mathcal{P}_\beta(q_{12},q_{13})$ and $\mathcal{P}_\beta(q_{12},q_{34})$ that read as

$$\mathcal{P}_\beta(q_{12},q_{13}) = \frac{1}{2}\mathcal{P}_\beta(q_{12})\mathcal{P}_\beta(q_{13}) + \frac{1}{2}\mathcal{P}_\beta(q_{12})\delta(q_{12}-q_{13}), \tag{1.8}$$

$$\mathcal{P}_\beta(q_{12},q_{34}) = \frac{2}{3}\mathcal{P}_\beta(q_{12})\mathcal{P}_\beta(q_{34}) + \frac{1}{3}\mathcal{P}_\beta(q_{12})\delta(q_{12}-q_{34}). \tag{1.9}$$

The first expression highlights that, when considering three replicas of the system, it turns out that either two of their overlaps are independent, or they are identical and these two outcomes happen with the same probability; the second expression confirms that, even when looking at overlaps between two distinct couples of replicas, hence considering four replicas, such a correlation persists although resized. Ultimately, this can be seen as a straight consequence of Parisi distribution for 3-replica overlaps that reads as

$$\begin{aligned}\mathcal{P}_\beta(q_1,q_2,q_3) = {}& \frac{1}{2}\mathcal{P}_\beta(q_1)x(q_1)\delta(q_1-q_2)\delta(q_1-q_3)\\ &+ [\mathcal{P}_\beta(q_1)\mathcal{P}_\beta(q_2)\theta(q_1-q_2)\delta(q_2-q_3) + \text{perm.}],\end{aligned} \tag{1.10}$$

where θ is the Heaviside function and $x(q)$ is the Parisi order parameter—for instance, if one marginalizes over q_3, gets $\mathcal{P}_\beta(q_1,q_2) = \frac{1}{2}\mathcal{P}_\beta(q_1)\delta(q_1-q_2) + \frac{1}{2}\mathcal{P}_\beta(q_1)\mathcal{P}_\beta(q_2)$, hence recovering (1.8) [71].

As a consequence of ultrametricity, along the past two decades a number of constraints on overlap fluctuations in the low temperature regime of spin glasses have been obtained in a mathematically controllable settings and, among these ensembles of families, the most famous ones are certainly the Ghirlanda–Guerra identities [42], whose simplest expressions read as

$$\langle q_{12}^4\rangle - 2\langle q_{12}^2 q_{13}^2\rangle + \langle q_{12}^2\rangle^2 = 0, \tag{1.11}$$

$$\langle q_{12}^4\rangle - 3\langle q_{12}^2 q_{34}^2\rangle + 2\langle q_{12}^2\rangle^2 = 0, \tag{1.12}$$

as well as their linear counterpart (where we get rid of $\langle q_{12}^2\rangle^2$ by substitution in the two equations above), obtained independently by Aizenman and Contucci [40] via stochastic stability (and later with several other techniques [41, 58, 59, 72]), whence the first identity of the family reads as

$$\langle q_{12}^4\rangle - 4\langle q_{12}^2 q_{23}^2\rangle + 3\langle q_{12}^2 q_{34}^2\rangle = 0. \tag{1.13}$$

Although the SK model remains the archetype of spin glasses, several variations on theme have appeared in the Literature, possibly relaxing its mean-field fully-connected nature. For instance, its version on random graphs (known as Viana–Bray model [43–45, 73]) was studied finding ultrametric fluctuations that naturally generalize Ghirlanda–Guerra and Aizenman–Contucci identities (and reduce to the latter whenever the coordination number of the graph approaches the network size). The same holds for models with higher-order interactions (known as P-spin models [46–48]), even in the diverging number of interactions (known as random energy model, REM [47]), up to extensions as neural networks (e.g. the Hopfield model) [52], and beyond [55–60, 74]. Further, more abstract representations of the SK model, as for instance the Random Overlap Structures introduced by Aizenman, Sims and Starr [49, 50] and its diluted RaMOSt counterpart, also exhibit Ghirlanda–Guerra fluctuations [51]. It is thus rather natural to further inspect the validity of these ultrametric constraints in glassy models of Natural Evolution, particularly focusing on *flat landscapes*.

2. Ultrametric fluctuations in glassy evolutionary models without selective pressure

Models such as Gardner's P-spin glass [75], Derrida's REM [76] or Hopfield's associative memory [23] have been shown to be plausible models for Natural Evolution under the presence of both random mutations and selective pressure (see e.g. [21, 22, 77]), also, they are well-known to exhibit overlap fluctuations that respect both the Ghirlanda–Guerra and the Aizenman–Contucci identities. However, moving to models of Natural Evolution taking place in flat landscapes nothing has been said so far on the validity of these ultrametric constraints. A possible difficulty in answering this question possibly lays in the absence of an Hamiltonian representation for these models. In the following we inspect the three best-known models in this context, that is the OPM (that is a model for asexual reproduction exhibiting, by construction, RSB on the scale of single progenies), the HPM (that is a basic model for sexual reproduction, where reproduction may involve two parents regardless their genetic distance and it is thus replica-symmetric) and the species formation model (SFM, that generalizes the previous model by requiring a threshold in string similarity for dating and this crucially turns the evolution of the model to be RSB at the level of species rather than single genomes).

We will show that, for the OPM, both the Ghirlanda-Guerra and the Aizenman-Contucci identities are violated and we prove the existence of another family of identities that is instead respected. Extending the same analysis on the HPM returns a rather simple scenario where all the identities are trivially respected (as anticipated since the model is replica-symmetric). Next, we tested (numerically) all the three families of ultrametric constraints on the SFM: a finite-size-scaling analysis suggests that they are expected to hold in the suitable limits (i.e. the infinite genome limit and large population limit), with the new class of identities being the ones minimally violated by the finite size effects. Driven by this last finding, we close this study by inspecting whether these constraints are fulfilled on actual genomes, focusing on a sample of the biological human genome and two artificial genomes and we find that the scenario depicted for the SFM is preserved also in these realistic settings.

The simplifying assumptions that we preserve along the paper are those of the original manuscripts (see e.g. [24]), namely

- while Evolution takes place, the population size is preserved and set equal to M;
- each individual $a \in \{1,\dots,M\}$ is represented by a string of N bits $\{S_1^a,\ S_2^a,\ \dots,S_N^a\}$, with N constant during the evolutionary process, which can be interpreted as the genome of the individual a[7];

[7] Actually, using a generic N-bits vectors allows us to map the string from a binary alphabet to the natural one for the problem under study (e.g. a quaternary one when dealing with the four DNA bases adenine (A), cytosine (C), guanine (G), and thymine (T)) such that, in general, the sequences $\{S_i^a(t)\}_{i=1}^N$ can represent bases of a nucleic acid sequence, amino acids in a protein, alleles in a genome, etc.

- the genome is subjected to mutations and we focus on point-mutations[8] that happen at constant mutation rate (along different generations) and independently of a given locus (i.e. the unit of the genome that mutates): we thus associate one-to-one to each genotype a phenotype[9].
- the dynamics is parallel: at each iteration all the individuals in the populations are removed and replaced by their offsprings.

With these simplifications the state of the population at a given time t can be described by specifying the genome of all the individuals $\{\boldsymbol{S}^a(t)\}_{a=1}^M$. The natural measure of genetic distance between two individuals a and b is the Hamming distance

$$d_{ab} = \frac{1}{2}\sum_{i=1}^{N} |S_i^a - S_i^b| = \frac{N}{2}(1 - q_{ab}), \tag{2.1}$$

where q_{ab} is the overlap between the genomes of the individuals a and b and mirrors the overlap between the spin configurations of two replicas. Analogously, the $M \times M$ matrix q evaluated at a given time t provides a snapshot of the population structure at that time. Interestingly, it can be proved that, in the $N \to \infty$ limit, the three flat landscape models under consideration can be simulated by directly looking at the evolution of q rather than dealing with the set of genomic sequences [36, 37].

2.1. The OPM

In the OPM studied by Derrida and Peliti [24] we consider a population Ω, made up of a fixed number, M, of individuals reproducing synchronously and asexually, whose genome at generation t is encoded by a N-bits vector $\boldsymbol{S}^a(t) \in \{-1, +1\}^N$ for $a = 1, \ldots, M$. At each generation t, all the individuals are removed, and a new generation is formed by offsprings of the previous individuals. More precisely, each individual $a \in \Omega$ is randomly associated to a parent $G(a) \in \Omega$ and the genome $\boldsymbol{S}^a(t)$ is taken identical to that of its parent $\boldsymbol{S}^{G(a)}(t-1)$ at the previous generation $t-1$ except for random mutations, as specified by

$$\mathbb{P}_1[S_i^a(t) = \pm S_i^{G(a)}(t-1)] = \frac{1}{2}(1 \pm e^{-2\mu}), \tag{2.2}$$

where $\mu \in \mathbb{R}^+$ tunes the mutation probability; the subscript '1' highlights that we are comparing individuals separated by one generation and, in the following, the expectation related to $\mathbb{P}_1$ shall be referred to as $\mathbb{E}_1$. As for the mapping $a \to G(a)$, it is assumed that $G(a)$ is chosen independently and uniformly in Ω for each individual and at each generation. Therefore, for any individual $a \in \Omega$, its ancestors over the previous t generations are given by the sequence $\{G(a), G^2(a), \ldots, G^t(a)\} = \Gamma_t(a)$.

As remarked in section 1, analogously to spin glasses, we have two averages: at each generation t we can take the average of any quantity involving the individuals of the whole population

[8] While real world mutation can include insertion and deletions [11] and more complex randomness, the theoretical advantage of single mutations is that a Markov process in the genome space driven by these mutation has symmetric transitions rates as if -say- genotype A is one-step away from genotype B, then also genotype B is one-step away from genotype A. Further, by the empirical counterpart there is a confirm [64] that the bulk of mutations in human genomes is point-like.

[9] In models with selective pressure the latter is used to evaluate the fitness of a given genotype such that the higher its fitness the larger the number of its offsprings, but this does not happen in flat landscapes.

Ω (*population average* $\langle\cdot\rangle$) but, as this quantity may fluctuate even for an infinitely large population Ω according to the particular mapping sequence (Γ_t) which has taken place, we should consider also the average of these quantities taken over all possible realizations of the reproduction process (*process average* $\overline{\cdot}$). Crucially, the process average can be obtained by averaging over the temporal unfolding of the process for a sufficiently-long time stretch. In fact, as shown in the next subsection, the typical overlap between the genomes of two individuals separated by Δt generations decays exponentially as $\exp(-\Delta t/\mathrm{const})$, with const $\propto M$. Thus, by taking a relatively-long time-span (we choose $10^2 \times M$), the time average is ensured to include a large sample of independent realizations.

Specifically, at generation t, the population average of the overlap, that we denote with $\langle q\rangle_t$, can be obtained by means of the following

$$\langle q\rangle_t = \int qP(q,t)\mathrm{d}q, \tag{2.3}$$

where

$$P(q,t) = \frac{1}{\binom{M}{2}}\sum_{a<b}\delta(q_{ab}(t)-q). \tag{2.4}$$

Thus, $\langle q\rangle_t$ fluctuates in time about a mean value that we denote with $\overline{\langle q\rangle}$ and which can be expressed as

$$\overline{\langle q\rangle} = \int q\bar{P}(q)\mathrm{d}q, \tag{2.5}$$

where $\bar{P}$ is the overlap distribution averaged over time. It can be proven [24] that, in the limit $N \gg 1$, the time-averaged overlap distribution depends only on the parameter $\lambda := \frac{1}{4M\mu}$ and it is

$$\bar{P}(q) = \begin{cases} \lambda q^{\lambda-1} & 0<q\leqslant 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

such that

$$\overline{\langle q\rangle} = \int q\bar{P}(q)\mathrm{d}q = \frac{\lambda}{\lambda+1}. \tag{2.7}$$

Notice that for $\lambda<1$ the distribution is peaked at $q=0$, for $\lambda=1$ the distribution is uniform in the interval $0<q\leqslant 1$, and as λ exceeds 1 the peaks is at $q=1$. As shown in figure 1, the agreement between theoretical predictions and simulation outcomes is pretty good already for relatively small sizes and it gets better and better as N is made larger. Remarkably, the broad distribution of the overlap q highlights the non-self-averaging nature of the order parameter in the OPM. Indeed, even in the infinite genome-size limit, one has [24]

$$\overline{\langle q\rangle^2} - \overline{\langle q\rangle}^2 \neq 0. \tag{2.8}$$

2.1.1. Exponential decay of correlations for efficient sampling.

If we let the system evolve for a time $t \gg 1$, the last generation will be made up of individuals with a unique common ancestor with probability one in the asymptotic limit [24], hence it is possible to find an expression for the decay of genome-correlations between individuals at a given time and their common ancestor, as a function of time. Let us start evaluating the expectation value of the overlap between a parent $S^{G(a)}(t)$ and the corresponding offspring at $t+1$:

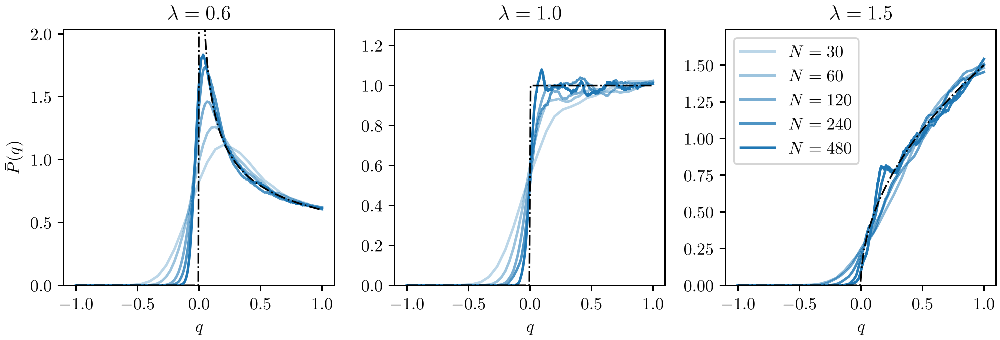

Figure 1. Numerical estimate for the overlap probability density function $\bar{P}(q)$ for a population of $M = 50$ individuals averaged over 10^5 generations. Several genome sizes are tested and depicted in different colors as explained in the legend, while the dashed curve corresponds to the theoretical probability distribution given by (2.6). As expected, if $\lambda < 1$ mutations are likely and the overlap concentrates on values $q \approx 0$, if $\lambda > 1$ mutations are rare and the overlap concentrates on values $q \approx 1$, while if $\lambda = 1$ the probability distribution becomes uniform.

$$q_{\Delta t=1} = \mathbb{E}_1\left[\frac{\boldsymbol{S}^{G(a)}(t)\cdot\boldsymbol{S}^a(t+1)}{N}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_1[S_i^{G(a)}(t)S_i^a(t+1)] \tag{2.9}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\{1 - 2\mathbb{P}_1[S_i^a(t+1) = -S_i^{G(a)}(t)]\} \tag{2.10}$$

$$= 1 - 2\mathbb{P}_1[S_i^a(t+1) = -S_i^{G(a)}(t)] = e^{-2\mu}, \tag{2.11}$$

where in the last line we exploited the fact that loci are independent. We can demonstrate that the expectation value of the overlap between a parent and the corresponding offspring at $t+\Delta t$ is

$$q_{\Delta t} = \mathbb{E}_{\Delta t}\left[\frac{S^{G^{\Delta t}(a)}(t)\cdot S^a(t+\Delta t)}{N}\right] = e^{-2\mu\Delta t}, \tag{2.12}$$

where the average $\mathbb{E}_{\Delta t}$ is performed over the distribution $\mathbb{P}_{\Delta t}$ which generalizes (2.2). We prove this by induction: first we observe that $q_{\Delta t=1} = e^{-2\mu}$ is true, next we assume that $q_{\Delta t} = e^{-2\mu\Delta t}$ is true for any Δt and we check that this is sufficient to ensure that also $q_{\Delta t+1} = e^{-2\mu(\Delta t+1)}$ is true. In the following, in order to lighten the notation, we shall set $t = 0$ without loss of generality and we shall drop the superscript labeling the individual without any loss of information: the individual we are referring to is a or its ancestor at the generation specified by time dependence of $\boldsymbol{S}$.

Let use observe that $q_{\Delta t}$ can be written as

$$q_{\Delta t} = \frac{1}{N}\sum_{i=1}\{1 - 2\mathbb{P}_{\Delta t}[S_i(\Delta t) = -S_i(0)]\} \tag{2.13}$$

and, analogously, $q_{\Delta t+1}$ can be written as

$$q_{\Delta t+1} = \frac{1}{N}\sum_{i=1}\{1 - 2\mathbb{P}_{\Delta t+1}[S_i(\Delta t+1) = -S_i(0)]\}\,. \tag{2.14}$$

By the law of total probability we can write

$$\begin{aligned}\mathbb{P}_{\Delta t+1}[S_i(\Delta t+1)=-S_i(0)] &= \mathbb{P}_{\Delta t}[S_i(\Delta t)=S_i(0)]\,\mathbb{P}_1[S_i(\Delta t+1)=-S_i(\Delta t)]\\ &\quad+\mathbb{P}_{\Delta t}[S_i(\Delta t)=-S_i(0)]\mathbb{P}_1[S_i(\Delta t+1)=S_i(\Delta t)]. \end{aligned} \tag{2.15}$$

Recalling that

$$\mathbb{P}_1[S_i(\Delta t+1)=-S_i(\Delta t)]=\frac{1-e^{-2\mu}}{2} \tag{2.16}$$

we reach

$$\mathbb{P}_{\Delta t+1}[S_i(\Delta t+1)=-S_i(0)]=\frac{1-e^{-2\mu}}{2}+e^{-2\mu}\mathbb{P}_{\Delta t}[S_i(\Delta t)=-S_i(0)]. \tag{2.17}$$

By direct substitution of the last equation into (2.14) we get

$$q_{\Delta t+1}=\frac{e^{-2\mu}}{N}\sum_{i=1}\{1-2\mathbb{P}_{\Delta t}[S_i(\Delta t)=-S_i(0)]\} \tag{2.18}$$

and, recalling the definition of $q_{\Delta t}$ given in (2.13), the last equation gets

$$q_{\Delta t+1}=\frac{e^{-2\mu}}{N}\sum_{i=1}[1-2\mathbb{P}_{\Delta t}(S_i(\Delta t)=-S_i(0))]=e^{-2\mu}q_{\Delta t}=e^{-2\mu(\Delta t+1)}. \tag{2.19}$$

In figure 2, the exponential decay of the overlap between the ancestor and its offsprings as a function of time is shown along with the related family tree.

2.1.2. A new class of ultrametric identities with three replicas. As stressed in section 2.1, in the OPM the genome overlap is non self-averaging and $\overline{\langle q^2\rangle}-\overline{\langle q\rangle}^2\neq 0$. Following spin-glass theory, one may wonder whether the overlap momenta are related by some non-trivial identities. As recalled in section 1.2, in the SK and many variations of its, the Ghirlanda–Guerra and Aizenman–Contucci identities are preserved under replica-symmetry-breaking. Here, to inspect the validity of these identities we study numerically the following quantities

$$\varepsilon_{\mathrm{GG}}=\overline{\langle q_{12}^4\rangle}-2\overline{\langle (q_{12}q_{13})^2\rangle}+\overline{\langle q_{12}^2\rangle^2} \tag{2.20}$$

$$\varepsilon_{\mathrm{AC}}=\overline{\langle q_{12}^4\rangle}-4\overline{\langle (q_{12}q_{13})^2\rangle}+3\overline{\langle q_{12}^2q_{34}^2\rangle} \tag{2.21}$$

$$\varepsilon_{\mathrm{SA}}=\overline{\langle q_{12}^4\rangle}-\overline{\langle q_{12}^2\rangle^2}, \tag{2.22}$$

where $\varepsilon_{\mathrm{GG,AC,SA}}$ are interpreted as a measure of possible violation. Remarkably, since our inspection is only based on numerics, at finite population size and along a finite time-span, in order to verify if non-null values of $\varepsilon_{\mathrm{GG,AC,SA}}$ are intrinsic or, rather, stem from finite-size effects, we will perform a finite-size-scaling: if the extent of ε is non-decreasing by increasing the system size, we will have a signature for the breakdown of the related identity.

Beyond these quantities, we can inspect the time-averaged joint probability density $\bar{P}(q_{12},q_{23},q_{13})$ of the 3-replica overlaps, in the infinite genome limit $N\to\infty$, that is known [24] and reads as,

$$\begin{cases}\bar{P}(q_{12},q_{23},q_{13})=\frac{\lambda^2}{2}\theta(q_{23}-q_{12})\delta(q_{12}-q_{13})q_{12}^{\lambda-1}q_{23}^{2\lambda-1}+\mathrm{Perm}(1,2,3) & q_{12},q_{13},q_{23}\in(0,1]\\ 0 & \text{otherwise}\end{cases} \tag{2.23}$$

looking for possible relations among overlap momenta involving three replicas.

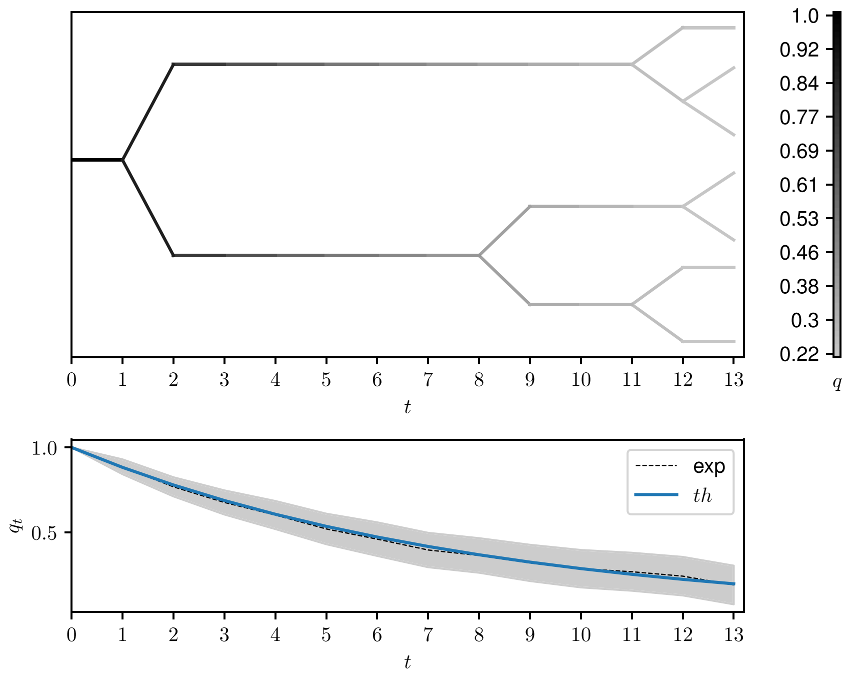

Figure 2. Upper panel: number of individuals generated by the ancestor as a function of the generation; in this particular evolution, all the individuals present at time $t = 13$ turn out to stem from the same common ancestor and other branches that have not survived are omitted in this plot. The colormap highlights the overlap between the various individuals and the ancestor. Lower panel: time decay of the overlap between the ancestor and its offsprings; the numerical estimate (solid line) is consistent with the theoretical estimate $e^{-2\mu t}$ (dashed line) obtained in (2.19); the shadow represents three standard deviations evaluated by repeating the process 100 times.

We stress that, at difference with the distribution provided in equation (1.10), in this scenario the replica symmetric solution (i.e. the first term in the r.h.s. in equation (1.10)) is never allowed: this behavioral difference between Parisi ultrametricity and the one pertaining to the OPM is expected to riverberate in slightly different ultrametric constraints as we deepen later. In particular, we find that in the thermodynamic limit $N \to \infty$

$$\overline{\langle q_{12}^K \rangle} = \frac{\overline{\langle q_{12} \rangle}}{K + \overline{\langle q_{12} \rangle}(1-K)} = \frac{\lambda}{\lambda + K}, \ \forall K \in \mathbb{N}, \tag{2.24}$$

which generalizes (2.7) and, by a direct calculation, we also find

$$\overline{\langle q_{12}^\alpha q_{13}^\alpha \rangle} = \frac{\lambda^2 \left(5\alpha^2 + 8\alpha\lambda + 3\lambda^2\right)}{(\alpha+\lambda)^2(2\alpha+\lambda)(2\alpha+3\lambda)}, \ \forall \alpha \in \mathbb{R}^+. \tag{2.25}$$

Now, by merging (2.24) and (2.25), we obtain the following relation, that plays as a new generator of overlap constraints for this model

$$\overline{\langle q_{12}^{2\alpha} \rangle} + \beta \overline{\langle q_{12}^\alpha q_{13}^\alpha \rangle} - (1+\beta) \frac{\overline{\langle q_{12}^\alpha \rangle}\,\overline{\langle q_{12}^{2\alpha} \rangle}\,\overline{\langle q_{12}^{2\alpha/3} \rangle}}{\overline{\left\langle q_{12}^{\frac{\alpha}{6}\left(5+\sqrt{\frac{25\beta+1}{\beta+1}}\right)} \right\rangle}\,\overline{\left\langle q_{12}^{\frac{\alpha}{6}\left(5-\sqrt{\frac{25\beta+1}{\beta+1}}\right)} \right\rangle}} = 0. \tag{2.26}$$

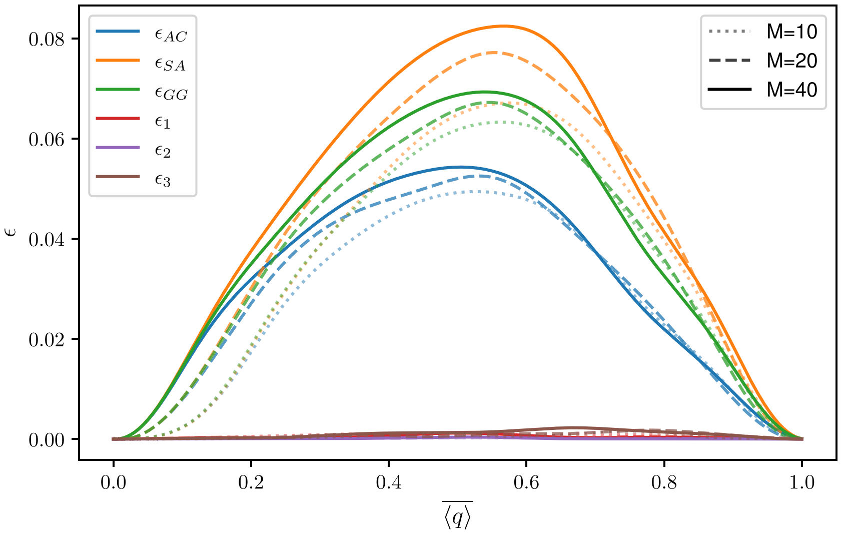

Figure 3. Errors measured for the various ultrametric identities (depicted in different colors, as explained by the legend on the left) as a function of the mean overlap evaluated for the OPM and for different values of M (depicted in different line styles, as explained by the legend on the right); the process average is taken over a run of $10^2\times M$ generations. Note that, as M (and the time span, accordingly) grows, the error on self-averaging ($\varepsilon_{\mathrm{SA}}$), on Ghirlanda-Guerra identities ($\varepsilon_{\mathrm{GG}}$) and on Aizenman-Contucci polynomials ($\varepsilon_{\mathrm{AC}}$) does not decrease, while the errors ε_1, ε_2, ε_3 are robustly vanishing.

Indeed, the above equation constitutes an infinite family of relations which hold for the OPM. In particular, due to the non-integrability of the overlap momenta with negative power, we must have $\beta\geqslant -1/25$ and $\alpha\geqslant 0$ which fulfills the condition $5-\sqrt{\frac{25\beta+1}{\beta+1}}\geqslant 0$ ensuring that the moments of the overlap are well defined.

As an example, if we set $\alpha=2$ and $\beta=1/7$ in equation (2.26) we get

$$\overline{\langle q_{12}^4\rangle}+\frac{1}{7}\overline{\langle q_{12}^2q_{13}^2\rangle}-\frac{8}{7}\frac{\overline{\langle q_{12}^2\rangle}\,\overline{\langle q_{12}^4\rangle}\,\overline{\langle q_{12}^{4/3}\rangle}}{\overline{\langle q_{12}^{7/3}\rangle}\,\overline{\langle q_{12}\rangle}}=:\varepsilon_1, \tag{2.27}$$

then, if we set $\alpha=2$ and $\beta=1/2$ in equation (2.26), we get

$$\overline{\langle q_{12}^4\rangle}+\frac{1}{2}\overline{\langle q_{12}^2q_{13}^2\rangle}-\frac{3}{2}\frac{\overline{\langle q_{12}^2\rangle}\,\overline{\langle q_{12}^4\rangle}\,\overline{\langle q_{12}^{4/3}\rangle}}{\overline{\langle q_{12}^{8/3}\rangle}\,\overline{\langle q_{12}^{\frac{2}{3}}\rangle}}=:\varepsilon_2 \tag{2.28}$$

where we introduced ε_1 and ε_2 to measure possible failures of these relations in analogy to (2.20)–(2.22). As shown in figure 3, both ε_1 and ε_2 are numerically found to vanish for the OPM. However, we stress that equations (2.27) and (2.28) are just two examples of equalities since there is an infinite family of relationships which are satisfied by the OPM and that can be obtained by varying $\alpha\in[0,+\infty)$ and $\beta\in[-1/25,+\infty)$ in equation (2.26). In figure 3 we also show numerical evidence that self-averaging is broken (as expected by construction) and that nor the Ghirlanda–Guerra identities neither the Aizenman–Contucci polynomials seem to hold.

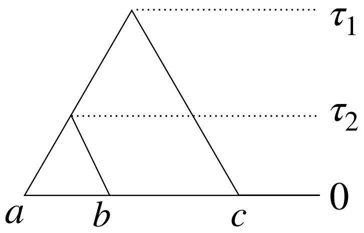

Figure 4. Diagrammatic representation of the only possible genealogy with three individuals with one common ancestor.

2.1.3. A new class of ultrametric identities with four replicas. In order to derive the equivalent information provided by the Ghirlanda–Guerra identities (see equations (1.11) and (1.13)) we need to account for four replicas.

Taken two individuals, labeled as a and b, we denote by τ_{ab} their *relatedness*, that is the rescaled number (rescaled w.r.t. M) of generations which we have to reckon back to find their latest common ancestor. Before evaluating the distribution of the relatedness between four different individuals let us introduce, following [24], the probability of a genealogy: given n individuals, the probability $\mathcal{G}_n\{\tau_{n-1},\dots,\tau_1\}$ of a genealogy with one common ancestor and with branching times (rescaled by M) $\tau_{n-1},\dots,\tau_1$ is

$$\mathcal{G}_n\{\tau_{n-1},\dots,\tau_1\} = \prod_{l=1}^{n} \exp\left[-\frac{l(l-1)}{2}(\tau_l - \tau_{l+1})\right] \tag{2.29}$$

with $\tau_n = 0$ and where l identifies the number of different lineages in the generations immediately following τ_l.

If $n = 3$ there is only one possible genealogy (see figure 4) with probability

$$\mathcal{G}_3\{\tau_1,\tau_2\} = \exp\left(-\tau_1 - 2\tau_2\right). \tag{2.30}$$

Then the probability that the individuals have relatedness $\bar{P}(\tau_{ab},\tau_{ad},\tau_{bd})$ can be constructed by steps: at first we have to multiply $\mathcal{G}_3\{\tau_1,\tau_2\}$ by the Heaviside step function to fix the temporal order of the branching times, that is

$$\theta(\tau_1 - \tau_2)\exp\left(-\tau_1 - 2\tau_2\right), \tag{2.31}$$

then by the delta function stemming from the fact that the relatedness of the individuals is set at the branching times (e.g. $\tau_1 = \tau_{ac} = \tau_{bc}$ and $\tau_2 = \tau_{ab}$), hence finally we get

$$\bar{P}(\tau_{ab},\tau_{ad},\tau_{bd}) = \frac{1}{2}\delta(\tau_{ac} - \tau_{bc})\exp\left(-\tau_{ac} - 2\tau_{ab}\right)\theta(\tau_{ac} - \tau_{ab}) + \mathrm{Perm}(a,b,c), \tag{2.32}$$

where $1/2$ is a normalization factor. By making the change of variable $q = e^{-\tau/\lambda}$ we get the overlap probability

$$\bar{P}(q_{ab},q_{ad},q_{bd}) = \frac{\lambda^2}{2}\delta(q_{ac} - q_{bc})\exp\left(-q_{ac} - 2q_{ab}\right)\theta(q_{ab} - q_{ac}) + \mathrm{Perm}(a,b,c). \tag{2.33}$$

Now, in order to evaluate the probability of the relatedness values between four individuals let us observe that there are three different ways in which we can add another individual to the three-individuals genealogy $\mathcal{G}_3\{\tau_1,\tau_2\}$ and they are represented by the dotted red lines in figure 5. The genealogies II and III are topologically equivalent, therefore the diagram II has multiplicity 2, ultimately providing

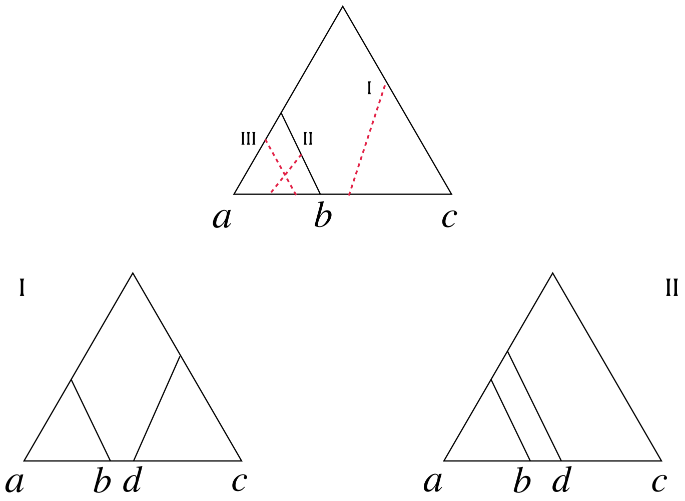

Figure 5. Genealogical trees obtained by adding to the genealogical tree with three individuals (upper side) another line accounting for the fourth individual.

$$\mathcal{G}_4\{\tau_1,\dots,\tau_3\} = \frac{1}{3}\mathcal{G}_4(\mathrm{I}) + \frac{2}{3}\mathcal{G}_4(\mathrm{II}), \tag{2.34}$$

where

$$\begin{aligned}\mathcal{G}_4(\mathrm{I}) = \frac{3}{4}&\theta(\tau_{dc}-\tau_{ab})\theta(\tau_{ad}-\tau_{dc})\exp\left(-\tau_{ad}-2\tau_{dc}-3\tau_{ab}\right)\\ &\times\delta\left(\tau_{ad}-\tau_{ac}\right)\delta\left(\tau_{ac}-\tau_{bd}\right)\delta\left(\tau_{bd}-\tau_{bc}\right)\end{aligned} \tag{2.35}$$

$$\begin{aligned}\mathcal{G}_4(\mathrm{II}) = \frac{3}{4}&\theta(\tau_{ac}-\tau_{ad})\theta(\tau_{ad}-\tau_{ab})\exp\left(-\tau_{ac}-2\tau_{ad}-3\tau_{ab}\right)\\ &\times\delta\left(\tau_{ad}-\tau_{bd}\right)\delta\left(\tau_{ac}-\tau_{bc}\right)\delta\left(\tau_{bc}-\tau_{dc}\right)\end{aligned} \tag{2.36}$$

and $3/4$ plays as a normalization factor. Then,

$$\begin{aligned}\bar{P}(\tau_{ab},\dots,\tau_{bd}) = \frac{1}{2}&\delta(\tau_{ad}-\tau_{bd})\delta(\tau_{bc}-\tau_{dc})\delta(\tau_{ac}-\tau_{bc})\theta(\tau_{ad}-\tau_{ab})\theta(\tau_{ac}-\tau_{ad})\\ &\times\exp\left(-3\tau_{ab}-2\tau_{ad}-\tau_{ac}\right)+\frac{1}{4}\delta(\tau_{ad}-\tau_{ac})\delta(\tau_{ac}-\tau_{bd})\\ &\times\delta(\tau_{bd}-\tau_{bc})\theta(\tau_{ad}-\tau_{cd})\theta(\tau_{dc}-\tau_{ab})\\ &\times\exp\left(-2\tau_{dc}-3\tau_{ab}-\tau_{ad}\right)+\mathrm{Perm}(a,b,c,d)\end{aligned} \tag{2.37}$$

By making the change of variable $q = e^{-\tau/\lambda}$ we get

$$\begin{aligned}\bar{P}(q_{ab},\dots,q_{bd}) = \frac{1}{2}&\lambda^3\delta(q_{ad}-q_{bd})\delta(q_{bc}-q_{dc})\delta(q_{ac}-q_{bc})\\ &\times\theta(q_{ab}-q_{ad})\theta(q_{ad}-q_{ac})q_{ab}^{3\lambda-1}q_{ad}^{2\lambda-1}q_{ac}^{\lambda-1}\\ &+\frac{1}{4}\lambda^3\delta(q_{ad}-q_{ac})\delta(q_{ac}-q_{bd})\delta(q_{bd}-q_{bc})\\ &\times\theta(q_{cd}-q_{ad})\theta(q_{ab}-q_{dc})q_{dc}^{2\lambda-1}q_{ab}^{3\lambda-1}q_{ad}^{\lambda-1}+\mathrm{Perm}(a,b,c,d).\end{aligned} \tag{2.38}$$

By exploiting the above overlap distribution it is possibile to evaluate overlap correlation functions involving four replicas as, e.g.

$$\overline{\langle q_{12}^{\alpha} q_{34}^{\alpha}\rangle} = \frac{\lambda^2\left(4\alpha^2+18\alpha\lambda+9\lambda^2\right)}{(\alpha+\lambda)(2\alpha+\lambda)(\alpha+3\lambda)(2\alpha+3\lambda)} = \frac{\overline{\left\langle q_{12}^{\frac{\alpha}{3}(3+\sqrt{5})}\right\rangle\left\langle q_{12}^{\frac{\alpha}{3}(3-\sqrt{5})}\right\rangle}}{\overline{\langle q_{12}^{\alpha}\rangle}\,\overline{\langle q_{12}^{2\alpha}\rangle}\,\overline{\left\langle q_{12}^{\frac{\alpha}{3}}\right\rangle}\,\overline{\left\langle q_{12}^{\frac{2\alpha}{3}}\right\rangle}}. \tag{2.39}$$

With some algebra, it has be proven that the following identity holds

$$\overline{\langle q_{12}^{2\alpha}\rangle} + \beta\overline{\langle q_{12}^{\alpha} q_{23}^{\alpha}\rangle} - (1+\beta)\overline{\langle q_{12}^{\alpha} q_{34}^{\alpha}\rangle}\frac{\overline{\left\langle q_{12}^{\frac{\alpha}{3}(3+\sqrt{5})}\right\rangle\left\langle q_{12}^{\frac{\alpha}{3}(3-\sqrt{5})}\right\rangle}}{\overline{\left\langle q_{12}^{\frac{\alpha}{6}\left(5+\sqrt{\frac{25\beta+1}{\beta+1}}\right)}\right\rangle\left\langle q_{12}^{\frac{\alpha}{6}\left(5-\sqrt{\frac{25\beta+1}{\beta+1}}\right)}\right\rangle}\,\overline{\left\langle q_{12}^{\frac{\alpha}{3}}\right\rangle}} = 0,$$

$$\forall\beta \geqslant -\frac{1}{25}, \alpha \geqslant 0, \tag{2.40}$$

providing the *equivalent information* in this new series of ultrametric constraints of the identity (1.12).

We can thus test the validity of these new constraints, considering also the possible violation of the above equation (that we evaluate for the case $\alpha = 2, \beta = 1/7$) introducing

$$\overline{\langle q_{12}^{4}\rangle} + \frac{1}{7}\overline{\langle q_{12}^{2} q_{23}^{2}\rangle} - \frac{8}{7}\overline{\langle q_{12}^{2} q_{34}^{2}\rangle}\frac{\overline{\left\langle q_{12}^{\frac{2}{3}(3+\sqrt{5})}\right\rangle\left\langle q_{12}^{\frac{2}{3}(3-\sqrt{5})}\right\rangle}}{\overline{\left\langle q_{12}^{\frac{1}{3}\left(5+\sqrt{\frac{25/7+1}{\frac{8}{7}}}\right)}\right\rangle\left\langle q_{12}^{\frac{1}{3}\left(5-\sqrt{\frac{25/7+1}{\frac{8}{7}}}\right)}\right\rangle}\,\overline{\left\langle q_{12}^{\frac{2}{3}}\right\rangle}} =: \varepsilon_3. \tag{2.41}$$

As shown in figure 3, ε_3 is numerically vanishing, similarly to what found previously for ε_1 and ε_2.

2.2. The HPM

Serva and Peliti [38] investigated a natural extension of the OPM, namely a two-parents model where parents can mate regardless their genome proximity; as a result of this feature, the long-time limit population is homogeneous (whence the name given to model) and, consequently, the model exhibits replica-symmetry in such a way that all the ultrametric constraints become trivial identities.

In the HPM, at each generation t, each individual a has two distinct parents $G_1(a)$ and $G_2(a)$ chosen at random from the previous generation. Each spin S_i^a is inherited from either $G_1(a)$ or $G_2(a)$ with equal probability, and probability of faithful copy or mutation as is the same as in equation (2.2). In the OPM model if the overlap between the parents $G(a)$ and $G(b)$ of two individuals, a and b, is $q_{G(a)G(b)}$ then the expectation value of the overlap of a and b is

$$q_{ab} = e^{-4\mu} q_{G(a)G(b)}. \tag{2.42}$$

If N is infinite this becomes a deterministic rule for updating the overlap matrix. There is an equivalent rule for updating the overlap matrix for the HPM in the limit $N \to \infty$. The pair of spins $S_i^a S_i^b$ is inherited from one of the four combinations of parents of the two individuals with equal probability, therefore

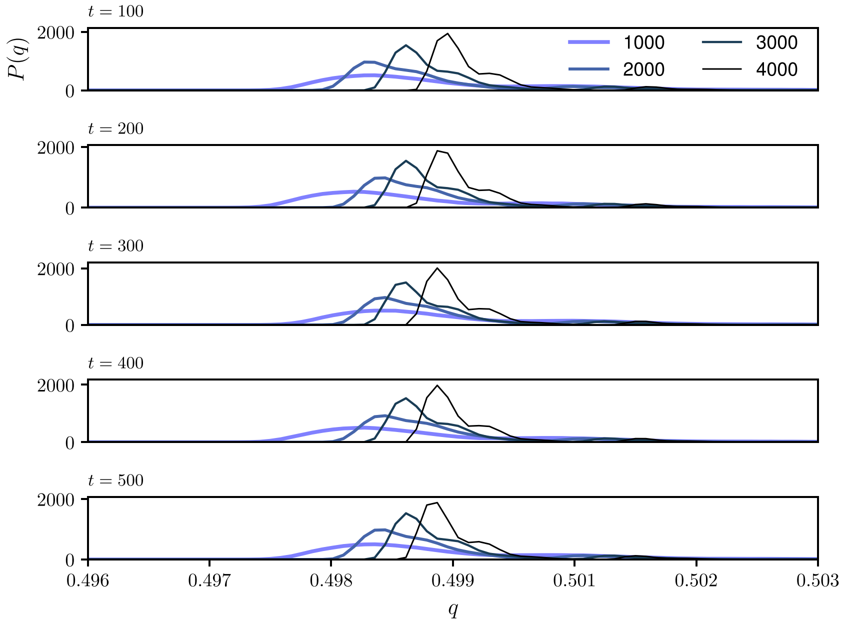

Figure 6. Distribution of the overlap $P(q,t)$ of the HPM at $\lambda = 1.0$ and for different values of M as shown in the legend (the thinner the line, the largest the size). As M is made larger and larger $P(q,t)$ gets monomodal and peaked at $\langle q\rangle_t = 1/2$, thus highlighting a replica symmetric behavior of the specie evolution in the HPM.

$$q_{ab} = \frac{e^{-4\mu}}{4}\left[q_{G_1(a)G_1(b)} + q_{G_2(a)G_1(b)} + q_{G_1(a)G_2(b)} + q_{G_2(a)G_2(b)}\right] \tag{2.43}$$

with $q_{aa} = 1$ always. It can be proven that the variance of $\langle q\rangle_t$ vanishes in the limit $M \to \infty$, thus $\langle q\rangle_t$ is self-averaging in the HPM, in particular $\lim_{M\to\infty} \overline{\langle q\rangle} = \frac{\lambda}{1+\lambda}$.

Figure 6 shows results about the overlap distribution for long simulations of the HPM model with $\lambda = 1$ and, accordingly, $\lim_{M\to\infty} \overline{\langle q\rangle} = 1/2$: the various rows show the overlap distribution $P(q,t)$ sampled at different times and by inspecting its variance as a function of M it can be shown that it scales as $1/M$ hence it is expected to disappear for large population size M such that $P(q,t)$ gets concentrated around $q = 1/2$, showing that self-averaging of the overlap is respected hence proving that the model behaves in a replica symmetric manner. Consistently, in figure 7 we show that the errors on the ultrametric identities (trivially) approach zero as M is made larger and larger.

2.3. The SFM

The SFM introduced by Higgs and Derrida [36, 37] is nothing but the two-parents model of Serva and Peliti with a threshold for mating $q_{\min}$. Specifically, it is defined in the same way as the HPM addressed in section 2.2, except that the first parent $G_1(a)$ of individual a is chosen randomly from the previous generation whereas the second $G_2(a)$ is chosen only from those individuals of the previous generation that display an overlap $q_{G_1(a)G_2(a)}$ greater

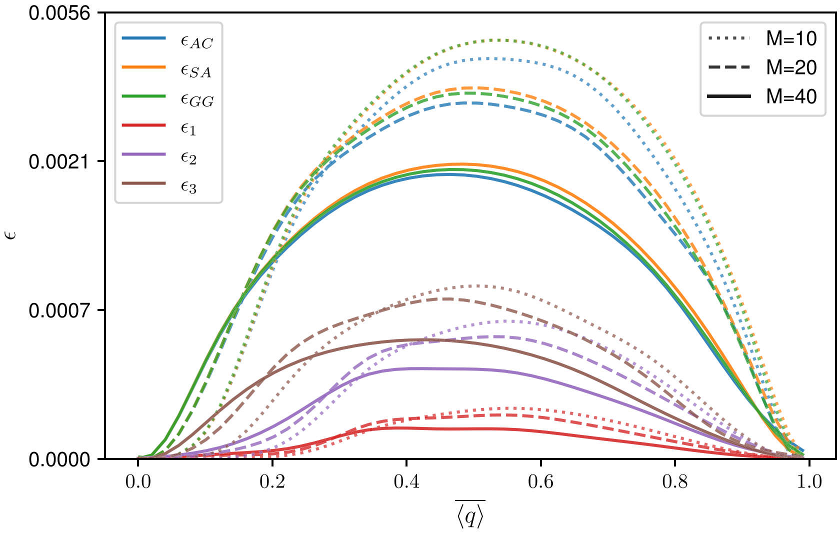

Figure 7. Errors measured for the various ultrametric identities as a function of the mean overlap for the HPM and for different values of M; the process average is taken over a run of $10^2 \times M$ generations Note that, as M grows, all the errors depicted tend to approach the horizontal axes as expected since the overlap is self-averaging in this model.

than a threshold value $q_{\min}$[10]. In the absence of a threshold, thus in the case of the HPM, there is a natural mean value of the overlap $\overline{\langle q \rangle} = \lambda/(1+\lambda)$, in such a way that, if we now set $q_{\min} > \lambda/(1+\lambda)$, the system is highly perturbed by the introduction of the threshold and it may never reach its natural equilibrium state: $\pi\acute{\alpha}\nu\tau\alpha\ \grave{\rho}\varepsilon\tilde{\iota}$ as in the low-temperature regime of spin glasses.

A corroboration of this picture appears in figure 8 where we show the distribution of the overlap $P(q,t)$ at different times: the peaks appearing above the threshold $q_{\min} > 0.65$ correspond to the overlaps of the new species that have formed (and their disappearance to the species extinction), while the large peak below the threshold exponentially collapses toward zero (since such values of the overlap are lower than $q_{\min}$ no interbreeding is possibile between the related genomes and, consequently, the peak must be vanishing).

2.3.1. Numerical inspection of ultrametric constraints in the SFM. We now study numerically the validity of all the ultrametric identities as well as of the self-averaging, by measuring the related errors ε. Specifically, we set $\lambda = 1$ in such a way that the expected value in the HPM is $\overline{\langle q \rangle} = \frac{1}{2}$ and we vary $q_{\min} \in [0,1]$. We simulate the evolution over a population of size M and a time span $10^2 \times M$, and we collect data for ε_1, ε_2, ε_3, $\varepsilon_{\mathrm{GG}}$, $\varepsilon_{\mathrm{SA}}$, $\varepsilon_{\mathrm{AC}}$ that are plotted in figure 9 versus $q_{\min}$. As expected, when $q_{\min} < \frac{1}{2}$, the threshold does not involve significant effects with respect to the HPM and a replica-symmetric scenario is recovered with all the errors ε close to zero. Conversely, the region $q_{\min} > \frac{1}{2}$ is non-trivial and there emerge differences between the

[10] If no such a second parent is available then a new first parent is randomly selected.

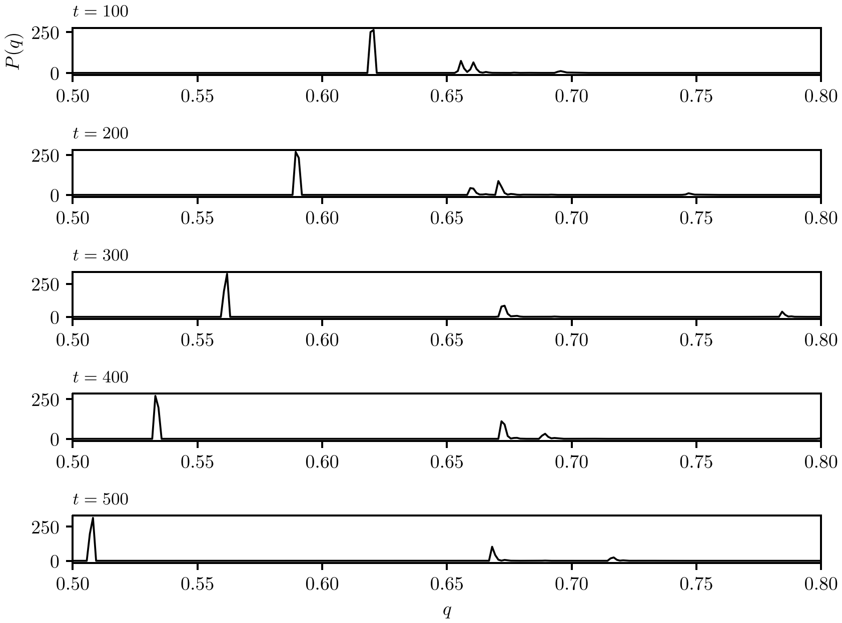

Figure 8. Distribution of the overlap $P(q,t)$ of the SFM for $\lambda = 1.0, q_{\min} = 0.65, M = 2000$. Note that, unlike the previous replica-symmetric HPM, here there actually are new specie formation and extinction as the presence of several peaks in the $P(q,t)$ evidences. Further, beyond the big one on the left (that is exponentially collapsing toward zero as time goes on), the erratic presence of these small peaks confirms that the SFM gives rise to an evolutive scenario strongly resembling the replica-symmetry-breaking of the low-temperature spin-glasses.

errors. As for the classical identities and for the variance (i.e. the self-averaging), the related errors ($\varepsilon_{\mathrm{GG,AC,SA}}$) are non-vanishing, actually their values grow with $q_{\min}$ without any robust trend with respect to the size M; as for the new identities, the related errors ($\varepsilon_{1,2,3}$) remain close to zero.

2.3.2. Numerical inspection of ultrametric constraints in biological and artificial human genomes. In this section we look for any evidence of the ultrametric relations discussed before in experimental genomic datasets. To this aim we tested all the ultrametric identities and the self-averaging property on the biological genome collected by *The 1000 Genomes Project Consortium* [64] and on artificial genomes generated by two neural networks (a Generative Adversarial network and a Restricted Boltzmann machine) [65] that have already proved to reproduce correctly allele frequencies, linkage disequilibrium, pairwise haplotype distances and population structure.

The *1000 Genomes Project Consortium* has undertaken a seven year research project (the 1*KGP*), set up in 2008, with the plan of providing high quality genomes out of 2504 people involving 26 populations across five continents, probably resulting in the most highly trustable collection of human genetic variations. Indeed the human genome comprises three billion

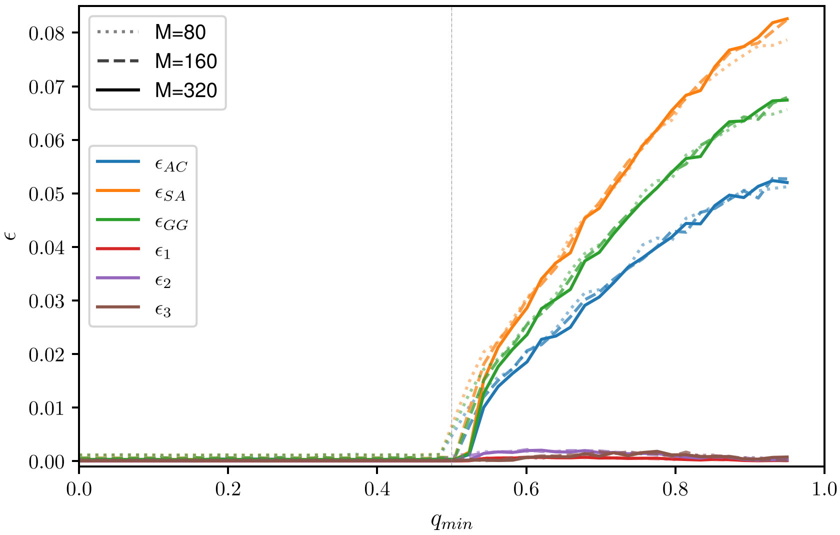

Figure 9. Errors measured for the various ultrametric identities as a function of the mating threshold q_{min} for the SFM and for different values of M; the process average is taken over a run of $10^2 \times M$ generations. At the numerical level, the new set of identities keep holding also for the SFM (i.e. $\varepsilon_1 \sim \varepsilon_2 \sim \varepsilon_3 \sim 0$), while all the other constraints seem to be violated.

bases (estimated to carry around 20 000 protein coding genes) and, given two random genomes, their average difference is of order of millions of bases[11], hence their careful sequencing is pivotal in providing highly informative repositories. The *1000 Genomes Project Consortium* has constituted a monumental challenge for bioinformaticians, who provided both whole-genome sequencing and targeted exome sequencing, and researchers have now validated 80 millions (out of 100) variants stored in the public database on SNP, hence it is the perfect target to test our new identities as well as the standard ones.

As in [65, 66], we consider a population of $M = 2504$ individuals ($\sim 5 \cdot 10^3$ haplotypes) spanning $N = 805$ Single Nucleotide Polymorphism (SNPs)[12] from [64], which reflect a high proportion of the population structure present in the whole dataset [65, 78]. The various fluctuation relations are evaluated by splitting the dataset of M individuals into $\sqrt{M}$ groups: the population average $\langle \cdot \rangle$ is carried out by identifying distinct replica indices with distinct individuals within the same group. In contrast, the process average $\bar{\cdot}$ is carried out by performing an arithmetic mean over the different evaluations of each group. Regarding the finite-size-scaling with N, it has been carried out by selecting a common subset of size N of the genome variable $(-1, +1)$ for each individual.

Results are collected in figure 10 and show that also in these structured datasets the new set of identities is better respected w.r.t. the classical ones (although the violation of the latter is minimal in these settings). Further, ε_3, that, we recall, is the error related to the new

[11] A typical genome differs from the reference human genome in 4–5 millions of sites and more than 99.9% of variants consist of SNPs on which the theoretical models discussed in this paper are built.

[12] Single nucleotide polymorphisms are the most common type of genetic variation among people: each SNP represents a difference in a single nucleotide (e.g. an SNP may replace the nucleotide cytosine C with the nucleotide thymine T in a certain stretch of DNA).

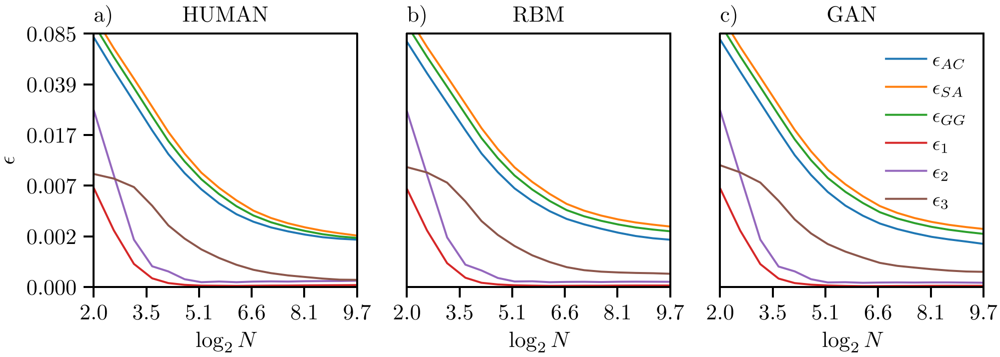

Figure 10. Finite-size-scaling on genome length testing the various ultrametric identities for the real human genome (HUMAN, taken from [64]) and two artificial genomes (taken from [65]) generated respectively with a Generative Adversarial Network (GAN) and a Restricted Boltzmann machine (RBM). In these settings still our new family of identities seems to be respected, while mild violations of the others persist (despite their violation is minimal, i.e. $\varepsilon_{\mathrm{AC}} \sim \varepsilon_{\mathrm{GG}} \sim O(10^{-3})$).

4-replica ultrametric identity, see equation (2.40), seems to be sensible to the origin of the dataset (biological vs artificial) as in the real biological dataset exhibits a faster drop as N is increased.

3. Conclusions

The non-self-averaging behavior of the order parameter in spin-glass models is a peculiar, intensively-studied feature which can be described in terms of a set of relations connecting the fluctuations of the order parameter. Driven by strong analogies between Natural Evolution and statistical mechanics of disordered systems, we investigated the validity of these ultrametric relations and the existence of other kinds of relations focusing on three stochastic models of evolving populations in flat landscapes. These models are the fairly standard ones in the Literature on Natural Evolution without selective pressure, that is (i) the OPM – where reproduction is asexual and the distribution of genetic distances lacks self-averaging – (ii) the HPM – where reproduction is sexual and with random mating (i.e. regardless the genetic distance) and thus results in a replica symmetric picture where the genetic distance between pairs of individuals has vanishing fluctuations in the thermodynamic limit—and (iii) the SFM where reproduction is still sexual but with a threshold on the required similarity among mating genomes before duplication. The latter represents the most interesting case as it is the closest to biology and it spontaneously gives rise to a complex dynamics reaching a steady state with new species that are continuously and spontaneously generated and suppressed during the evolutionary process. Further, while in the first model the evolutionary tree is assumed and it is related to single descendants from common ancestors, in the latter the evolutionary tree emerges and it works at the level of species rather than single elements.

Focusing on fluctuations in the genetic distances between individuals, as far as the OPM is concerned, after checking that self-averaging is absent in this model statistics, we have shown by a finite-size-scaling argument that nor the Ghirlanda–Guerra identities neither the Aizenman–Contucci polynomials are respected. On the other hand, we were able to prove a new class of identities that are indeed respected also in our finite-size numerical checks. For the HPM, as it is replica symmetric, all these constraints are equally guarantee to converge to zero in the asymptotic limit but they do not convey actual information. Then, dealing with the SFM, our identities continue to hold, being only mildly affected by finite-size effects.

As a final test we focused on human genomes: we considered the real biological dataset taken from the *1000 genome project consortium* and two synthetic datasets on artificial genomes generated by neural networks and, for all these three cases the scenario depicted by the SFM seems to be confirmed here as well: the new set of ultrametric identities is sharply respected while mild violations affect both Ghirlanda–Guerra identities as well as Aizenman–Contucci polynomials. Clearly, the validity of these new constraints should be made statistically robust with more and more analysis on real datasets, and it is difficult to state by now their role in a near future, but possible practical applications could involve shortening the timeline of sequence analysis, result in novel filtering techniques to clean raw dataset and more.

Further, as models of Natural Evolution under selective pressure (namely Darwinian Evolution) are known to display standard Ghirlanda–Guerra fluctuations (see e.g. the Franz-Peliti-Sellitto model, corresponding to the $P \to \infty$ limit of the Kauffman–Levin P-spin-glass model or the REM in the spin-glass jargon, and the equal-trap model analyzed by Leuthäusser and Tarazona corresponding to the Hopfield model in the spin-glass jargon), a similar analysis to the present one should be conducted also in these *not flat landscape* scenarios to better understand the role covered by Natural Selection (beyond random mutation) in shaping evolutionary taxonomies because, at present, these new findings seem to be in better agreement with Kimura Theory of Neutral Evolution.

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

This work is supported by Ministero degli Affari Esteri e della Cooperazione Internazionale (MAECI) through the grant *BULBUL* (F85F21006230001), and by The Alan Turing Institute through the Theory and Methods Challenge Fortnights event ‘Physics-informed machine learning’, which took place on 16-27 January 2023 at The Alan Turing Institute headquarters. The authors acknowledge financial support from PNRR MUR Project PE0000013-FAIR, from PRIN MUR Project 20229T9EAT and from Sapienza University of Rome (RM120172B8066CB0).
The authors are indebted with Pierluigi Contucci, Aurélien Decelle, Silvio Franz, Francesco Guerra, Luca Peliti and Beatriz Seoane for countless precious discussions.

ORCID iD

Elena Agliari https://orcid.org/0000-0002-5121-3511

References

[1] Parisi G 1980 A sequence of approximated solutions to the SK model for spin glasses *J. Phys. A: Math. Gen.* **13** L115

[2] Parisi G 1980 The order parameter for spin glasses: a function on the interval 0–1 *J. Phys. A: Math. Gen.* **13** 1101

[3] Van Mourik J and Coolen A C C 2001 Cluster derivation of Parisi's RSB solution for disordered systems *J. Phys. A: Math. Gen.* **34** L111

[4] Amit D J 1989 *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press)

[5] Tkacik G, Marre O, Mora T, Amodei D, Berry I I M J and Bialek W 2013 The simplest maximum entropy model for collective behavior in a neural network *J. Stat. Mech.* **2013** 03011

[6] Huang H 2021 *Statistical Mechanics of Neural Networks* (Springer)

[7] Braunstein A, Pagnani A, Weigt M and Zecchina R 2008 Inference algorithms for gene networks: a statistical mechanics analysis *J. Stat. Mech.* 12001

[8] Torrisi G, Kuehn R and Annibale A 2020 Percolation on the gene regulatory network *J. Stat. Mech.* 083501

[9] Mora T, Walczak A M, Bialek W and Callan C G 2010 Maximum entropy models for antibody diversity *Proc. Natl Acad. Sci.* **107** 5405–10

[10] Agliari E, Barra A, Del Ferraro G, Guerra F and Tantari D 2015 Anergy in self-directed B lymphocytes: a statistical mechanics perspective *J. Theor. Biol.* **375** 21–31

[11] Murugan A, Mora T, Walczak A M and Callan C G 2012 Statistical inference of the generation probability of T-cell receptors from sequence repertoires *Proc. Natl Acad. Sci.* **109** 16161–6

[12] Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M and Walczak A M 2012 Statistical mechanics for natural flocks of birds *Proc. Natl Acad. Sci. USA* **109** 4786–91

[13] Solé R V and Bascompte J 2006 *Self-Organization in Complex Ecosystems* (Princeton University. Press)

[14] Solé R V, Manrubia S C, Benton M, Kauffman S and Bak P 1999 Criticality and scaling in evolutionary ecology *Trends Ecol. Evol.* **14** 156–60

[15] Agliari E, Barra A, Contucci P, Pizzoferrato A and Vernia C 2018 Social interaction effects on immigrant integration *Palgrave Commun.* **4** 55

[16] Durlauf S N 1999 How can statistical mechanics contribute to social science? *Proc. Natl Acad. Sci. USA* **96** 10582–4

[17] Coolen A C C 2005 *The Mathematical Theory of Minority Games: Statistical Mechanics of Interacting Agents* (Oxford University Press)

[18] Cristelli M, Gabrielli A, Tacchella A, Caldarelli G and Pietronero L 2013 Measuring the intangibles: a metrics for the economic complexity of countries and products *PLoS One* **8** e70726

[19] Mézard M, Parisi G and Zecchina R 2002 Analytic and algorithmic solution of random satisfiability problems *Science* **297** 812–5

[20] Mézard M and Montanari A 2009 *Information, Physics and Computation* (Oxford University Press)

[21] Kauffman S and Levin S 1987 Towards a general theory of adaptive walks on rugged landscapes *J. Theor. Biol.* **128** 11–45

[22] Franz S, Peliti L and Sellitto M 1993 An evolutionary version of the random energy model *J. Phys. A: Math. Gen.* **26** L1195

[23] Tarazona P 1992 Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin-glass models *Phys. Rev.* A **45** 6038

[24] Derrida B and Peliti L 1991 Evolution in a flat fitness landscape *Bull. Math. Biol.* **53** 355–82

[25] Eigen M 1993 Viral quasispecies *Sci. Am.* **269** 42–49

[26] Nowak M and May R M 2000 *Virus Dynamics: Mathematical Principles of Immunology and Virology* (Oxford University Press)

[27] Agliari E, Barra A, Burioni R, Di Biasio A and Uguzzoni G 2013 Collective Behaviours: from biochemical kinetics to electronic circuits *Sci. Rep.* **3** 3458

[28] Agliari E, Altavilla M, Barra A, Dello Schiavo L and Katz E 2015 Notes on stochastic (bio)-logical gates: computing with allosteric cooperativity *Sci. Rep.* **5** 9415

[29] Agliari E, Barra A, Dello Schiavo L and Moro A 2016 Complete integrability of information processing by biochemical reactions *Sci. Rep.* **6** 36314

[30] Svensson E and Calsbeek R (eds) 2012 *The Adaptive Landscape in Evolutionary Biology* (Oxford University Press)

[31] Kimura M 1983 *The Neutral Theory of Molecular Evolution* (Cambridge University Press)

[32] Felsenstein J 1981 Evolution ary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates *Evolution* **35** 1229–42

[33] Lake J A 1994 Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances *Proc. Natl Acad. Sci. USA* **91** 1455–9

[34] Frank S A and Slatkin M 1992 Fisher's fundamental theorem of natural selection *Trends Ecol. Evol.* **7** 92–95

[35] Leuthäusser I 1987 Statistical mechanics of Eigen's evolution model *J. Stat. Phys.* **48** 343–60

[36] Higgs P and Derrida B 1991 Stochastic models for species formation in evolving populations *J. Phys. A: Math. Gen.* **24** L985

[37] Higgs P and Derrida B 1992 Genetic distance and species formation in evolving populations *J. Mol. Evol.* **35** 454–65

[38] Serva M and Peliti L 1991 A statistical model of an evolving population with sexual reproduction *J. Phys. A: Math. Gen.* **24** L705

[39] Peliti L 1997 Introduction to the statistical theory of darwinian evolution (arXiv:cond-mat/9712027)

[40] Aizenman M and Contucci P 1998 On the stability of the quenched state in mean-field spin-glass models *J. Stat. Phys.* **92** 765–83

[41] Contucci P and Giardiná C 2007 The Ghirlanda-Guerra identities *J. Stat. Phys.* **126** 917–31

[42] Ghirlanda S and Guerra F 1998 General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity *J. Phys. A: Math. Gen.* **31** 9149

[43] Barra A and De Sanctis L 2007 Stability properties and probability distributions of multi-overlaps in dilute spin glasses *J. Stat. Mach.* 08025

[44] De Sanctis L and Franz S 2007 Self-averaging identities for random spin systems *Spin Glasses: Statics and Dynamics: Summer School* vol 62 ed A Bovier and A B Monvel (birkhäuser)

[45] Franz S and Leone M 2003 Replica bounds for optimization problems and diluted spin systems *J. Stat. Phys.* **111** 535–64

[46] Burioni R, Barra A, Burioni R and Di Biasio A 2012 Notes on the p-spin glass studied via Hamilton-Jacobi and smooth-cavity techniques *J. Math. Phys.* **53** 063304

[47] Contucci P and Giardiná C 2013 *Perspectives on Spin Glasses* (Cambridge University Press)

[48] Panchenko D 2010 The Ghirlanda Guerra identities for mixed p-spin model *Comp. Ren. Math.* **348** 189–92

[49] Aizenman M, Sims M R and Starr S L 2003 Extended variational principle for the Sherrington-Kirkpatrick spin-glass model *Phys. Rev.* B **68** 214403

[50] Chen W K 2013 The Aizenman-Sims-Starr scheme and Parisi formula for mixed p-spin spherical models *Electr. Journ. Prob.* **18** 1–14

[51] Sollich P and Barra A 2012 Spin glass polynomial identities from entropic constraints *J. Phys. A: Math. Theor.* **45** 485001

[52] Barra A and Guerra F 2009 Constraints for the order parameters in analogical neural networks (arXiv:0911.3113)

[53] Bovier A and Kurkova I 2004 Derrida's generalized random energy models 2: models with continuous hierarchies *Ann. H.P. Prob. and Stat.* **40** 481–95

[54] Bovier A, Kurkova I 2003 Derrida's generalized random energy models 4: continuous state branching and coalescents No. MP-ARC-2003-247 WIAS

[55] Contucci P and Giardiná C 2005 Spin-glass stochastic stability: a rigorous proof *Ann. H. Poincare* **6** 5

[56] Contucci P, Giardiná C and Giberti C 2009 Interaction-flip identities in spin glasses *J. Stat. Phys.* 1181–203

[57] Chatterjee S 2009 The Ghirlanda-Guerra identities without averaging (arXiv:0911.4520)

[58] Arguin L P 2008 Competing particle systems and the Ghirlanda-Guerra identities *Electron. J. Probab.* **13** 2101–17

[59] Chen Y T 2019 Universality of Ghirlanda-Guerra identities and spin distributions in mixed p-spin models *Ann. Inst. H. Poincaré Prob. Stat.* **55** 528–50

[60] Talagrand M 2011 The ghirlanda-guerra identities *Mean Field Models for Spin Glasses* vol II pp 287–311

[61] Panchenko D 2010 A connection between the Ghirlanda-Guerra identities and ultrametricity *Ann. Prob.* **38** 327–47

[62] Panchenko D 2011 Ghirldanda-Guerra identities and ultrametricity: an elementary proof in the discrete case *C. R. Math.* **349** 813–6

[63] Panchenko D 2013 The Parisi ultrametricity conjecture *Ann. Math.* **177** 383–93

[64] 2015 The 1000 genomes project consortium, a global reference for human genetic variation *Nature* **526** 68–74

[65] Yelmen B, Decelle A, Ongaro L, Marnetto D, Tallec C, Montinaro F, Furtlehner C, Pagani L and Jay F 2021 Creating artificial human genomes using generative neural networks *PLoS Gen.* **17** e1009303

[66] Decelle A, Rosset L and Seoane B 2023 Unsupervised hierarchical clustering using the learning dynamics of RBMs *Phys. Rev. E* **108** 014110

[67] Sherrington D and Kirkpatrick S 1975 Solvable model of a spin-glass *Phys. Rev. Lett.* **35** 1792

[68] Dotsenko V S 1993 Physics of the spin-glass state *Phys.-Usp.* **36** 455

[69] Guerra F 2003 Broken replica symmetry bounds in the mean field spin glass model *Comm. Math. Phys.* **233** 1–12

[70] Talagrand M 2006 The Parisi formula *Ann. Math.* **163** 221–63

[71] Rammal R, Toulouse G and Virasoro M A 1986 Ultrametricity for physicists *Rev. Mod. Phys.* **58** 765

[72] Barra A 2006 Irreducible free energy expansion and overlaps locking in mean field spin glasses *J. Stat. Phys.* 601–14

[73] Viana L and Bray A J 1985 Phase diagrams for dilute spin glasses *J. Phys.* C **18** 3037

[74] Contucci P, Giardiná C, Giberti C, Parisi G and Vernia C 2007 Ultrametricity in the Edwards-Anderson model *Phys. Rev. Lett.* **99** 057206

[75] Gardner E 1985 Spin glasses with p-spin interactions *Nucl. Phys.* B **257** 747–65

[76] Derrida B 1980 Random-energy model: limit of a family of disordered models *Phys. Rev. Lett.* **45** 79–82

[77] Peliti L 2002 Quasispecies evolution in general mean-field landscapes *Europhys. Lett.* **57** 745

[78] Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y and Tyler-Smith C 2014 Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences *Genome Biol.* **15** R88