PLOS Complex Systems

REVIEW

Component systems: Do null models explain everything?

Andrea Mazzolini[1,2,3], **Mattia Corigliano**[4,5], **Rossana Droghetti**[4,5], **Matteo Osella**[1,2], **Marco Cosentino Lagomarsino**[4,5,6]*

1 Dipartimento di Fisica, Università di Torino, Turin, Italy, **2** I.N.F.N., Turin, Italy, **3** I.N.F.N., Bari, Italy, **4** IFOM ETS, The AIRC Institute of Molecular Oncology, Milan, Italy, **5** Dipartimento di Fisica, Università degli Studi di Milano, Milan, Italy, **6** I.N.F.N., Milano, Italy

* marco.cosentino-lagomarsino@ifom.eu

OPEN ACCESS

Citation: Mazzolini A, Corigliano M, Droghetti R, Osella M, Cosentino Lagomarsino M (2026) Component systems: Do null models explain everything? PLOS Complex Syst 3(6): e0000114. https://doi.org/10.1371/journal.pcsy.0000114

Editor: Chiara Poletto, University of Padova: Universita degli Studi di Padova, ITALY

Published: June 30, 2026

Funding: This work was supported by the Italian "Ministero dell'Università e della Ricerca", PRIN 2022 – COD. 2022PY8MHN – GeCoS: Genomic Component Systems (MCL, MO). MC and RD received salary from these funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Component systems — ensembles of realizations built from a shared repertoire of modular parts — are ubiquitous in biological, ecological, technological, and socio-cultural domains. From genomes to texts, cities, and software, these systems exhibit statistical regularities that often meet the *bona fide* requirements of laws in the physical sciences. Here, we argue that the generality and simplicity of those laws are often due to basic combinatorial or sampling constraints, raising the question of whether such patterns are actually revealing system-specific mechanisms and how we might move beyond them. To this end, we first present a unifying mathematical framework, which allows us to compare modular systems in different fields and highlights the common "null" trends as well as the system-specific uniqueness, which, arguably, are signatures of the underlying generative dynamics. Next, we can exploit the framework with statistical mechanics and modern machine-learning tools for a twofold objective. (i) Explaining why the general regularities emerge, highlighting the constraints between them and the general principles at their origins, and (ii) "subtracting" them from data, which will isolate the informative features for inferring hidden system-specific generative processes, mechanistic and causal aspects.

Introduction

The study of complex systems has emerged as a central theme in modern science, driven by the recognition that many phenomena, from biological systems to technological infrastructures, share underlying organizational principles. A particularly productive framework for understanding these systems is the concept of modularity. Modularity refers to the decomposition of a complex system into smaller, relatively independent components that interact to produce system-level behavior [1]. This perspective on empirical data is not new; it has been applied in fields ranging from biology [2,3] to software engineering [4] and even urban planning [5]. In the last decades, advances in data availability, computational power, and theoretical tools have opened

new avenues for exploring modularity in unprecedented detail, as well as for understanding its origins [6–8].

Across domains as diverse as evolutionary genomics, linguistics, software engineering, and urban studies, complex systems can often be represented as realizations assembled from modular components, or "component systems". Genomes consist of gene families; texts consist of words; microbiomes are assemblies of individuals from different taxa; LEGO kits consist of plastic bricks; single-cell transcriptomes consist of expressed genes; and even food recipes are composed of ingredients [9]. Despite their heterogeneity, these systems reveal common empirical regularities. Well-known examples include Zipf's law for component frequencies [10–17], the sublinear power-law-like scaling of diversity -or vocabulary size- with system size (known as Heaps' law in linguistics) [16,18–20], and the widespread U-shaped occurrence patterns of gene families in genomics [16,21,22].

These patterns have been discovered independently in several fields, and system-specific generative processes have been proposed to explain their emergence. The common mathematical framework proposed in this Review can help the recognition of the generality of these patterns and promote the fruitful exchange of ideas and data analysis techniques between fields.

The component system and its emerging laws

A unifying description of component systems reveals that their apparent complexity is governed by a small set of robust and reproducible statistical regularities. These patterns, often described as emerging "laws", recur across systems [16,19,23–29] with widely differing microscopic details. Before entering the details of these results, we first introduce a common notation.

A unifying notation for component systems

The basic feature of a modular or component system is that its *realizations* — genomes, books, LEGO sets — are made of elementary *components* — genes, words, LEGO bricks — that can be reused within and between realizations (Fig 1(a)). This can be encoded in a matrix of integer elements n_{ij}, indicating how many times the component i appears in the realization j (Fig 1(b)).

Starting from this matrix, we introduce the first set of definitions involving summations of rows or columns (Fig 1(c)). The sum within a realization, $m_j = \sum_i n_{ij}$, defines its *size* in terms of how many components it contains, and it can be used to normalize the counts defining the frequency $f_{ij} = n_{ij}/m_j$ (Table 1). The sum across realizations of a given component, $a_i = \sum_j n_{ij}$, is the *abundance* of that component in the ensemble. The component abundances are often normalized to obtain an ensemble frequency $g_i = a_i / \sum_{i'} a_{i'}$.

A next set of informative quantities (Table 1) involves the binarized matrix that describes the presence-absence patterns. First we have the *sharing number* or *occurrence* of a component that counts in how many realizations the component is present, $o_i = \sum_j \Theta(n_{ij} - 1)$, where $\Theta(x)$ is the Heaviside function, which is 1 for $x \geq 0$. Its transposed quantity is the *vocabulary* of a realization, indicating the number its different components $h_j = \sum_i \Theta(n_{ij} - 1)$.

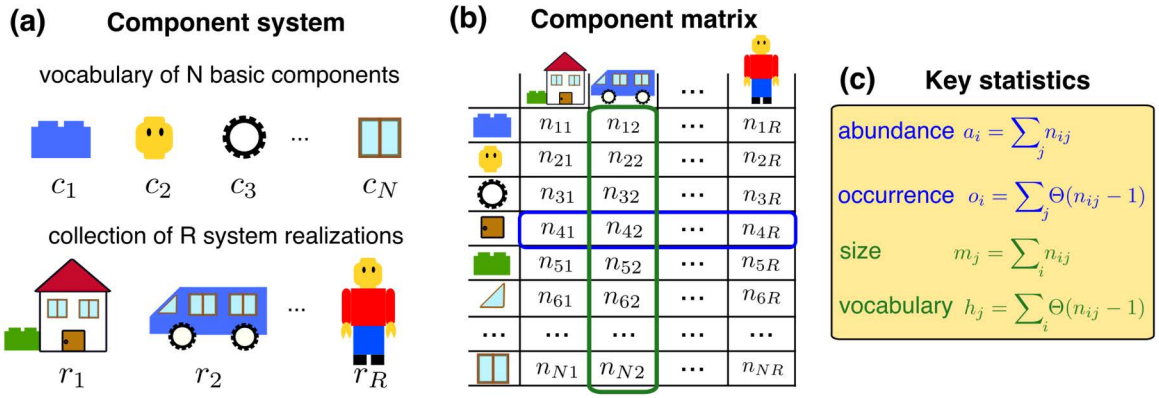

Fig 1. Component systems as a unifying representation of modular artifacts. (a) A wide class of systems—including genomes, texts, and LEGO constructions—can be described as *component systems*, in which each realization is assembled from a shared vocabulary of elementary components that may be reused within and across realizations. **(b)** This structure is encoded by the component matrix n_{ij}, whose entries count how many times component *i* appears in realization ***j***. **(c)** Simple sums and binarized sums of the matrix elements define a set of fundamental observables that characterize the system at both the component and realization level, including component abundance and occurrence, as well as realization size and vocabulary.

https://doi.org/10.1371/journal.pcsy.0000114.g001

Table 1. List of main mathematical symbols of a component system.

n_{ij}	**count of component *i* in realization *j***
f_{ij}	frequency of component *i* in realization *j*
m_j	size of realization *j*
h_j	vocabulary of realization *j*
a_i	abundance of component *i*
g_i	ensemble frequency of component *i*
o_i	occurrence/sharing-number of component *i*

https://doi.org/10.1371/journal.pcsy.0000114.t001

At this level of description, the component system is a weighted bipartite network [30] with an adjacency matrix n_{ij}, where abundance and size correspond to the node degrees on the two layers. This parallel allows us to benefit from many tools that have been introduced in network theory. For example, measures of node centrality that define the "importance" of a component or realization [31,32], or community detection and topic modeling techniques to find groups of realizations characterized by similar component compositions [33–37]. However, we argue that a component system deserves a specific notation for two main reasons. The first is conceptual: the class of systems that we are describing here have a precise interpretation of the two layers — realizations and components — and meaning of the introduced quantities — abundance, size, sharing number and vocabulary —, that would be lost with the network notation. The second reason is the fact that a component system has an additional set of properties described below that go beyond bipartite networks.

Components can often be grouped into classes that play similar roles in determining the functioning of a realization. For instance, protein-coding genes may be classified according to structural or evolutionary criteria, or into coarser functional families, while words can be grouped into syntactic or semantic categories. This hierarchical organization enables the definition of macro-components and allows the same system to be described at multiple levels of resolution. Such multiscale structure is a defining feature of component systems and introduces additional constraints and observables that are not naturally captured by standard bipartite network representations. At the same time, realizations can have a natural hierarchical structure, as genomes can be grouped into taxonomical units, books into different topics and LEGOs into different

themes. As discussed later, the study of how the system statistics change by coarse-graining at the level of components and realizations is still under-explored.

Closely related to this multiscale organization is the presence of dependencies among components, which constrain their joint occurrence within realizations [38]. In many systems, a given module can appear only if other modules are present, as exemplified by software dependencies in Linux distributions [39]. More generally, one can posit the existence of an (often hidden) dependency network among components, whose non-trivial structure gives rise to functional clusters that naturally correspond to the macro-components introduced above.

A final additional structure, present only in some component systems, is the temporal ordering of components. This ordering is naturally defined in linguistic artifacts, where words follow a well-defined sequence, while it is more blurred or even inaccessible in other systems. For instance, in LEGO constructions the order of bricks may be associated with the sequence of assembly steps, yet multiple construction paths can lead to the same final composition, and such information is rarely available in public datasets. The challenge is even more pronounced in genomic systems, where ordering may reflect evolutionary times of gene acquisition, but its precise definition and empirical reconstruction remain elusive. Nevertheless, when a meaningful notion of order can be identified, it enables the definition of additional observables and statistics that encode valuable information about the system's organization and generative mechanisms [20,40,41].

The emerging "physical laws" of component systems

The definition of a common language between different systems opens up the possibility of exporting concepts and results between fields. In particular, the study of emerging patterns has a long tradition in linguistics [42] and ecology [28,43], and can now be systematically extended to many other systems. The fascination of these patterns or laws is due to their mathematical simplicity and robustness, while the underlying systems are complex and heterogeneous. In this regard, extending the range of systems in which those regularities can be comparatively studied provides more hints about general and robust features or, instead, system-specific behaviors [16,27]. Before going to the critical discussion about the information these "laws" carry and how to exploit them to better understand a system, below we present a list of the more studied patterns.

Zipf, Heaps, and component-sharing laws. A first class of empirical regularities concerns statistics obtained from the row and column sums of the component matrix n_{ij}, as well as of its binarized counterpart (see Table 1). Among these patterns, the most extensively documented is Zipf's law, originally identified in linguistics [10] and later observed across a wide range of systems [11,27,44,45]. Zipf's law describes the distribution of component frequencies, either at the ensemble level through g_i or, depending on the context, within individual realizations through f_{ij}.

Defining the rank r_i of a component as its position in the list of components ordered by decreasing frequency (with $r_i = 1$ denoting the most frequent component), Zipf's law states that $g_i \propto r_i^{-\mu}$, where the exponent μ is typically close to unity (Fig 2(a)). The extensive literature devoted to this regularity, as well as to the various mechanisms proposed to explain it, is reviewed elsewhere [17,44,46].

Shifting the focus from component-level to realization-level statistics naturally leads to Heaps' law, originally introduced in the context of quantitative linguistics [18,47,48], and subsequently extended to other domains, including genomics [19], innovation dynamics [49,50], and diversity scaling in socioeconomic and biological systems [51]. Heaps' law characterizes how the vocabulary size h_j of a realization grows with its total size m_j, capturing the progressive slowdown in the introduction of new components as realizations become larger. Empirically, the average vocabulary size at a given realization size follows a sublinear scaling, $\langle h \rangle \propto m^{\nu}$, with $0 < \nu < 1$ (Fig 2(c)). Recent studies have also addressed the fluctuations around this average scaling behavior [27,52,53], revealing a robust quadratic scaling of the variance, $\sigma_h^2 \propto \langle h \rangle^2$, in agreement with Taylor's law [54].

A third class of regularities concerns the sharing statistics of components, quantified by the occurrence o_i, defined as the fraction of realizations in which a given component is present. The distribution of o_i has been extensively studied in

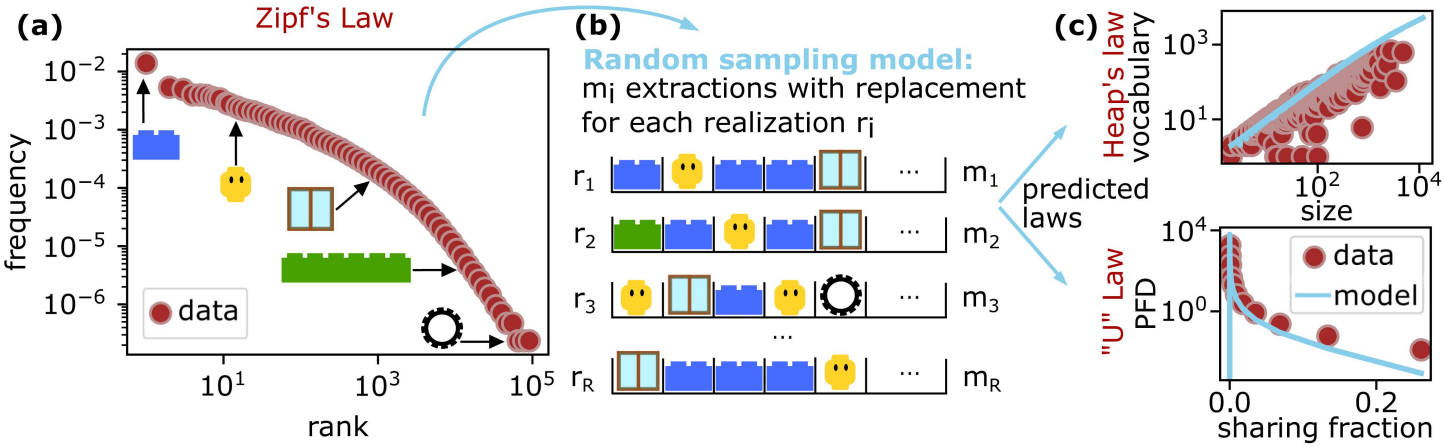

Fig 2. Basic statistical laws and their connection due to sampling constraints. (a) Zipf's law for component frequency is reported for the LEGO dataset (https://rebrickable.com/). **(b)** By extracting components with their empirical probabilities, a random sampling model can generate an artificial ensemble of realizations that can be compared to the empirical ones. **(c)** The statistics of shared components and the Heaps' law can be often explained by this sampling procedure, although quantitative deviations can reveal system-specific mechanisms.

https://doi.org/10.1371/journal.pcsy.0000114.g002

comparative genomics [21,22,55], where it exhibits a characteristic U-shaped form. This structure reflects the coexistence of a small set of highly conserved "core" components, present in nearly all realizations ($o_i \simeq 1$), alongside a large number of "accessory" or rare components that appear in only a small fraction of realizations ($o_i \sim 1/R$, where R denotes the total number of realizations). A related example arises in quantitative immunology, where the sharing of T-cell receptor clones across individuals is analyzed. In this context, the occurrence distribution has been used to test and validate generative models of receptor sequences, whose generation probabilities are strongly heterogeneous. This result holds in "public" receptors, which are easier to produce and likely shared between many individuals, as opposed to "private" clones, which tend to be unique to only one person [56,57].

Beyond the statistics of row and column summations. Statistics that go beyond simple row and column summations of the component matrix typically lack the universality and robustness of the aforementioned laws, and often display substantial heterogeneity both within and across component systems. For this reason, they are comparatively less prominent in the literature, despite being potentially more informative about system-specific organization and underlying generative mechanisms.

In genomics, for example, Grilli and coworkers [58] analyzed the statistics of the frequencies f_{ij} by fixing a given gene family i and studying its distribution across genomes. They found that the resulting distribution shapes vary widely among gene families, and linked this variability to differing evolutionary processes, including rates of horizontal gene transfer and family-specific expansion dynamics.

Instead of comparing frequencies of components across realizations, one can also consider their rank r_{ij}, i.e., the position in the list of ordered f_{ij} in a given realization j. Even if the overall frequency statistics robustly follows a power law across realizations, the ranking of components can change in a non-trivial way [59]. For example, comparison of rank-orders between realizations can reveal specific trends, especially in the changes among rare words, with common words being more stable [60]. Different studies have also focused on the rank dynamics by comparing realizations at different points in time, providing insights on the mechanistic temporal changes of the system in natural, social, economic, and infrastructural systems [61,62].

Additionally, different studies, especially in transcriptomics (the study of gene expression profiles across cells, environments and conditions), have investigated geometrical properties of the set of genes expressed in different types of cells,

claiming that the effective manifold in which cells function is low-dimensional [63,64], and this dimension decreases as cells progressively differentiate [65,66]. This concept is also present in immunology, where the host immune memories and the viral strains are often assumed to interact and evolve in a low-dimensional space, leading to consistent predictions of observed patterns [67,68]. The geometrical properties of this interaction was also analyzed in comparative experiments where the binding of different antibodies was tested against panels of different flu strains [69–71].

In linguistics, a substantial body of work has explored more refined statistical observables [42]. One example, which may be relevant to other component systems, is the tendency of rare words at the ensemble level to cluster within specific documents while being absent from most others [72], a behavior that deviates from null models based on independent sampling from ensemble frequencies. Additional approaches focus on sequential properties, such as the entropy of word sequences [73], or more generally on statistics that explicitly account for temporal order [40]. A variety of statistical and probabilistic models have been proposed to reproduce these patterns and to serve as suitable null generative models for data analysis and mining, including Poisson-based and bursty-process frameworks [74,75], but the debate remains open regarding the selection among different candidate scenarios.

Including functional categories. In principle, all the discussed statistics can be studied by coarse-graining the matrix of counts through *a priori* grouping of components into functional categories or by grouping realizations based on similarity. There are examples of studies of this type in microbial ecology [76], where specific marginal statistics can be explored by varying the phylogenetic similarity between species.

In genomics, different works have studied the scaling relation between the amount of gene families belonging to a macro-category, e.g., genes involved in metabolism or transcription factors, and the genome size [23,77,78]. The scaling is a power-law-like growth with a category-dependent exponent, which can be understood in terms of basic assembly rules that depend on the category function [26,79,80]. Crucially, this behavior cannot be obtained by sampling models, and is not found in most other empirical component systems [16,78]. Interestingly, the same type of nonlinear category-specific scaling has been also reported for socio-economic units within cities, e.g., transportation, scientific units, agricultural activities and many more [29], showing, again, a category dependent growth as the city size increases. In that context, different activity classes were argued to follow sublinear, linear, or superlinear scaling, reflecting their position within an urban functional hierarchy, an interpretation that is remarkably similar to the ones proposed for the genomic counterpart [26,79]. For urban data, the observed saturation of diversity in the largest cities was interpreted as finite classification resolution, rather than from an intrinsic limit to diversification, pointing to an underlying open-ended growth of functional diversity with system size.

Key regularities emerge from sampling constraints

A central outcome of recent work on component systems, ranging from genomes to software, is the identification of robust null trends that emerge even in the absence of specific biological or functional constraints. In particular, equilibrium sampling models have been shown to reproduce several empirical regularities, taking as input a fixed frequency distribution, in close analogy with ensembles of random networks with a prescribed degree sequence [81,82].

The simplest multinomial sampling model

In its simplest formulation, the sampling probabilities are taken to be equal to the empirical Zipf's law frequencies g_i, and each realization is composed of a number of independently drawn components equal to its size m_j (Fig 2(b)). This defines a multinomial process that generates a new set of component counts $\{n'_{1j}, \ldots, n'_{Nj}\}$ sampled from a multinomial distribution with m_j trials and $\{g_1, \ldots, g_N\}$ probabilities. One possibility is to use the exact empirical marginals, therefore generating a new artificial copy with minimal addition of assumptions. The second is to assume a specific analytical function for one marginal, i.e., a power law frequency distribution, providing a powerful tool to intuitively understand marginal parameter dependencies, but usually being valid only approximately and in a regime of large-sizes.

Despite its minimal assumptions, this framework provides a natural explanation for the sublinear growth observed in Heaps' law (Fig 2(c)), predicting an average scaling $\langle h \rangle \propto m^{1/\mu}$, where μ denotes the Zipf exponent [13,83,84]. The same null model also yields quantitative predictions for the statistics of shared components across realizations. Taken together, these results demonstrate that strong connections between different marginal observables—such as Zipf's law, Heaps' law, and component sharing—can arise purely from sampling effects. Empirically, a wide range of component systems are found to approximately follow these predictions (Fig 2(c)) [16], indicating that information is often redundant across laws. This redundancy suggests that different regularities, as well as different systems, should be analyzed jointly rather than in isolation.

Going beyond the multinomial model. The sampling model above is just an example of statistical-mechanics models that can describe the consequences of a given input hypothesis, in this case the Zipf's law for the universe of components, on the different statistical laws of the system. Alternatives can be based on maximum entropy ensembles that fix the degree distribution or other desired statistical properties, which find already application in classical network theory [85]. Another direction is to add constraints where components carry intrinsic limits, cross-enrichment or self-enrichment in their abundance. Models of this kind can explain features such as universal low-abundance yet ubiquitous gene families—can arise from occupancy [58,86]. Similarly, partition statistics derived from "least constrained" ensembles recapitulate macro-level scaling patterns across functional categories [23] and provide baseline expectations for condensation-like phenomena, where a small number of components dominate system-wide usage [87,88]. These results highlight that a surprisingly large portion of large-scale regularities emerge from a heterogeneous component abundance usage, described by the Zipf's law, in combination with sampling and combinatorial structure alone, before introducing any explicit evolutionary or functional coupling.

Finite-size effects and sampling. Empirical component systems are necessarily observed at finite size, which can affect the estimation of statistical laws such as rank-frequency relations, occurrence statistics, and related scaling exponents. In particular, deviations from asymptotic behavior are most pronounced in the low-frequency regime, where undersampling can distort both tails of the distribution and derived quantities such as Heaps' law or occurrence curves. These effects are naturally incorporated when the empirical marginal distributions are used as input to the sampling model, since finite-size variability is then propagated consistently to all derived observables [38]. For other generative models finite-size corrections may introduce systematic biases, and care is required in interpreting scaling relations. A theoretically interesting example is provided by the Pitman-Yor (Chinese Restaurant) duplication-innovation process, where finite-size effects and lack of self-averaging can lead to realization-dependent distortions of Zipf scaling, including anomalous cutoffs and nontrivial condensation-like structure in the occupation distribution that persist up to large system sizes [88]. Sampling-based null models provide a baseline that explicitly accounts for finite-size effects and allows one to distinguish robust structural features from artifacts at fixed sampling depth.

Subtracting regularities unveils system-specific mechanistic and causal processes

But if these sampling models account for so many large-scale patterns, across such different systems, does this make them irrelevant, or worse, trivial? And how can we then go beyond them? It is first important to stress that sampling models are built upon assumptions that typically include Zipf's law, which carries system-specific signatures in some contexts and for some questions. For example, in texts, deviations from the observed exponent can be a sign of corrupted/inefficient communication [89], or they can be used to define and characterize extreme events [90]. Equally, in genomes, Zipf's law is considered a positive signature of evolution by gene duplication, and the variance of gene-family abundance distributions across genomes correlates with the rate of horizontal transfers [2,25,58]. However, in most cases, a power-law-like abundance distribution can be generated by several alternative mechanisms, making it by itself unsuitable to discriminate or falsify generative models. One prominent example is found in ecology, where the species-abundance statistics can be generated by both of the two main competing scenarios that were proposed [91], a neutral assembly of

species and niche-structured communities. In such cases, all the other marginal statistics generated by sampling models will not provide additional information.

A similar debate exists for gene-family occurrence statistics across genomes (the fraction of families occurring in a given fraction of genomes). As we discussed, this statistics follows a U-shaped pattern, which has been interpreted as a signature of enriched classes of gene families performing core versus system-specific functions as a result of selection [25,55]. However, null/neutral sampling models can also give rise to U-shaped gene-frequency distributions [92], and a close comparison with a sampling model shows that null trends and enrichment coexist in the data, and only the families associated to key genome information processes (DNA replication, transcription and translation) emerge as a true core beyond null trends [16].

Following this route, we propose that studying such systems by employing suitable sampling models as null/neutral models can disentangle precisely positive and null trends. We intend here "null/neutral" models in a more general sense than used in statistics, in the sense that, beyond performing statistical tests, studying their general behavior can be useful to describe conceptually connections between marginal statistics, what information they contain and propose useful observables and indicators. This leads to two complementary outcomes. First, successful sampling predictions across systems reveal what is common and unifying across modular systems — from genomes to software —– by identifying the statistical signatures that arise from key general ingredients and constraints, related to combinatorics, generative rules (e.g., use of module copy-pasting [19,93]), statistical dependencies, and system size. This allows us to identify and develop an intuition regarding the general behavior shared by diverse architectures, providing a baseline notion of universality. Second, once this baseline is established, we can identify and infer the system-specific mechanisms that deviate from the null expectations: dependency hierarchies, evolutionary constraints, functional specialization, correlated innovations, and ecological pressures. These models act as a magnifying lens, allowing principled inference of finer structure that would otherwise be obscured by the overwhelming general trends.

A simple proposal to quantify such deviations is based on the Z-score. If we consider a generic observable of the count matrix $A(\{n_{ij}\})$, we can compare its value against K independent instances of the multinomial sampling process (or the most suitable null model). We can compute A on each of the the k-th instance of the sampling, $\{n'^{(k)}_{ij}\}$, finding its average $\mu_A = \sum_k A(\{n'^{(k)}_{ij}\})/K$ and standard deviation $\sigma_A^2 = \sum_k \left(A(\{n'^{(k)}_{ij}\}) - \mu_A\right)^2/(K-1)$. The deviation of the empirical observation from the expected outcome of the null model in units of standard deviations is then given by $Z_A = \left(A(\{n_{ij}\}) - \mu_A\right)/\sigma_A$. This number can then separate cases of redundancy with the frequency statistics and sampling constraints, $Z_A \sim 0$, and observables carrying new information with $|Z_A|$ larger than a chosen threshold.

More generally, against this clarified baseline, key specific model ingredients, for example dependency structures and history-dependent novelty processes [12,19,38,49] introduce deviations in properly chosen observables that match empirical signatures across systems. For example, dependency-structure based models treat genomes or software systems as directed acyclic graphs in which components require other components to be functional [38]. Imposing these asymmetries alters both the occurrence distributions and their scaling with system size, yielding patterns that cannot be reproduced by equilibrium sampling alone [39]. These enriched models produce key testable predictions linking module observables to underlying dependency hierarchies.

When null models are coupled with inference frameworks, deviations themselves become informative. For genomes, deviations in family-size scaling can indicate evolutionary potentials, dependency structures, or horizontal gene transfer. In software, they may reveal architectural constraints, robustness bottlenecks, or organizational principles. In texts, deviations from Zipf or Heaps laws reflect semantic or syntactic organization.

A comparative science of component systems offers a unifying lens on different domains. Texts, for instance, can be treated as assemblies of component words whose statistical patterns—embodied in Heaps' and Zipf's laws—show the balance between innovation and reuse as well as deeper semantic dependencies. LEGO kits offer a complementary engineered setting in which component diversity, design constraints, and creativity can be quantified [16]. In biological

contexts, microbiomes present functional gene repertoires as components whose organization follows macroecological laws observable in metagenomic data [28]. Single-cell RNA-seq provides a parallel cellular perspective, with genes acting as components whose usage varies across cells [27,35] and ultimately defines the cell identity [65,94]. Together, these arenas enable robust tests of the generality of statistical-mechanical approaches and help reveal which features of component systems—growth, dependency, innovation, or constraints—drive both their shared signatures and their domain-specific differences.

A plethora of theoretical models have been proposed to explain the origin of one or more of the statistical laws we discussed. From stochastic processes based on realization growth by duplication and innovation dynamics — including processes based on the sample-space dynamics or on correlated novelty processes [20,49,95,96] - to more abstract statistical mechanics descriptions based on the presence of latent or unobserved variables [15,97,98].

Finally, these considerations also open the way for model-selection and Bayesian inference approaches, which allow to compare model performance within general settings. Indeed, a clear understanding of the formal connections or the crucial differences between these alternative theoretical descriptions is still missing. Statistical properties of empirical component systems that go beyond the basic laws we described (such as fluctuation properties or correlation patterns) can represent a new testing ground for model selection.

Conclusions and perspectives

Taken together, these results suggest a general strategy for the study of component systems. Robust large-scale regularities emerge from minimal assumptions about sampling, combinatorics, and growth, defining a baseline notion of universality across domains. Against this baseline, systematic deviations—whether driven by dependencies, functional constraints, or history-dependent innovation—become the primary carriers of mechanistic and causal information. This perspective shifts the focus from an observational catalog of empirical laws to a quantitative framework that uses equilibrium models, such as the multinomial sampling, to quantify relationships between them and how much information they carry. These models are not directly used to learn the generative process, but as tools to identify the best set of observables for then comparing generative mechanisms, inferring hidden structure, and identifying system-specific trends.

Notably, Zipf-like distributions, the basic ingredient on which random sampling models are built, may emerge in multiple contexts, including machine learning, from the inference process in under-sampled, high-dimensional conditions [98,99]. These studies show that Zipf's like behaviour and other features of criticality in data can arise based on the trade-off between "relevance", the entropy of the frequency distribution of a given set of labels (e.g., the occurrence of gene families), and "resolution", the entropy of the label distribution, measuring the level of detail at which data are represented (e.g., by a classification into gene families), which can be properties of the data and not of the system under study. We implicitly assumed here that in component systems similar distributions arise already at the level of the system itself, as a consequence of heterogeneous components and intrinsic modular structure. This may be very likely for some systems (e.g., LEGO sets) but it is less clear for others (e.g., texts). In other words, Zipf's law may be a hybrid phenomenon: partly a reflection of the system's intrinsic modular organization and partly a consequence of strong under-sampling and optimal inference or learning dynamics. In the latter case, the observation of a Zipf law should be interpreted as a signature of the relevance of the measured variables, rather than a direct indicator of any particular system behavior, but the joint marginal laws would still be observable in data. Disentangling these contributions for specific empirical systems will be informative on their nature.

Despite progress, significant challenges remain. A key theoretical gap is a unified statistical mechanical theory that classifies the diverse generative models for component systems, bridging equilibrium sampling, history-dependent processes, and rule-based assembly. Both theoretically and empirically we lack systematic methods to studying the granularity ("coarse-graining") of components across scales. It remains unclear whether, how, and why marginal distributions and statistical patterns transform across hierarchical levels of description, and how this affects the identification of universal

behavior and its dependence on scale. Advances along these lines would also clarify the consequences of adopting different definitions of elementary components, which are often partly arbitrary. For example, sequence similarity criteria in genomics or taxonomic assignment methods in microbial ecology rely on chosen thresholds that need to be systematically analyzed. Equally, the implications of component-level laws for larger-scale phenomena, such as for example the interplay between genome composition and microbiome function or resilience, remain largely uncharted territory.

This perspective also raises a forward-looking methodological question related to modern machine learning and complex structured data: given a modular data structure, how can baseline statistical regularities and system-specific deviations be learned and characterized in increasingly complex, high-dimensional systems?

To address this question, we propose that Large Language Models (LLMs) trained on large-scale datasets of language, code, and biological sequences represent a promising avenue. Indeed, LLMs can be seen as generators of token sequences subject to learned complex constraints, so that the word sequences generated by LLMs not only reproduce classical statistical regularities such as Zipf's and Heaps' laws [100,101], but also capture complex patterns that go well beyond word frequencies [102]. Therefore, LLM-generated sequences represent a suitable setting to explore and understand structured complex data.

Although a well-defined statistical-mechanical description of these models is still lacking, the higher-order statistical regularities they capture likely extend the empirical laws discussed in this perspective, many of which remain to be characterized. Token-token correlations, temporal correlations, and higher-order interactions are likely crucial to the ability of LLMs to generate high-quality text and code. A full quantitative characterization of these structural patterns, together with minimal mathematical null models able to reproduce them, remains an open problem.

Finally, analyzing how statistical regularities vary with context length, temperature, model size, and across different models may provide insight into the underlying generative mechanisms, as well as into the emergence of transitions in the ability to capture increasingly complex patterns. Such transitions may be related to phenomena such as grokking—a sudden shift from memorization to generalization [103]—and to the acquisition of complex capabilities [104]. At present, these questions remain largely unexplored and represent a promising direction for future research.

To sum up, the central theme of this perspective is that *null models* based on sampling, diversification, and innovation provide a shared language to compare disparate component systems. At the same time, systematic deviations from these null trends reveal the key processes -biological, cognitive, technological, ecological etc.- that make each system unique. Component systems provide a unifying lens for studying complexity across disciplines. By combining null models, inference tools and insights from modern machine learning, we can build a coherent statistical mechanics of component systems. Such a framework has the potential to clarify the origins of observed regularities, reveal the mechanisms that shape complex architectures, and guide data-driven discovery in several scientific domains.

Acknowledgments

We thank Miguel Muñoz, Marco Gherardi, James Holehouse, Yasser Roudi, and Jacopo Grilli for useful discussions and suggestions.

Author contributions

Conceptualization: Andrea Mazzolini, Matteo Osella, Marco Cosentino Lagomarsino.

Visualization: Mattia Corigliano, Rossana Droghetti.

Writing – original draft: Marco Cosentino Lagomarsino.

Writing – review & editing: Andrea Mazzolini, Mattia Corigliano, Rossana Droghetti, Matteo Osella, Marco Cosentino Lagomarsino.

References

1. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A. 2006;103(23):8577–82. https://doi.org/10.1073/pnas.0601602103 PMID: 16723398
2. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002;420(6912):218–23. https://doi.org/10.1038/nature01256 PMID: 12432406
3. Lapenta F, Jerala R. Design of novel protein building modules and modular architectures. Curr Opin Struct Biol. 2020;63:90–6. https://doi.org/10.1016/j.sbi.2020.04.006 PMID: 32505942
4. Koch S. Software evolution in open source projects—a large-scale investigation. J Softw Maint Evol: Res Pract. 2007;19(6):361–82. https://doi.org/10.1002/smr.348
5. Barthélemy M. Spatial networks. Physics Reports. 2011;499(1–3):1–101. https://doi.org/10.1016/j.physrep.2010.11.002
6. Kashtan N, Alon U. Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci U S A. 2005;102(39):13773–8. https://doi.org/10.1073/pnas.0503610102 PMID: 16174729
7. Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. Nat Rev Genet. 2007;8(12):921–31. https://doi.org/10.1038/nrg2267 PMID: 18007649
8. Lorenz DM, Jeng A, Deem MW. The emergence of modularity in biological systems. Phys Life Rev. 2011;8(2):129–60. https://doi.org/10.1016/j.plrev.2011.02.003 PMID: 21353651
9. Caprioli C, Kulkarni S, Battiston F, Iacopini I, Santoro A, Latora V. The networks of ingredient combinations as culinary fingerprints of world cuisines. NPJ Sci Food. 2025;9(1):242. https://doi.org/10.1038/s41538-025-00588-4 PMID: 41266345
10. Zipf GK. Human behavior and the principle of least effort. 1949.
11. Li W. Zipf's Law everywhere. Glottometrics. 2002;5(2002):14–21.
12. Corominas-Murtra B, Solé RV. Universality of Zipf's law. Phys Rev E Stat Nonlin Soft Matter Phys. 2010;82(1 Pt 1):011102. https://doi.org/10.1103/PhysRevE.82.011102 PMID: 20866560
13. Lü L, Zhang Z-K, Zhou T. Zipf's law leads to Heaps' law: analyzing their relation in finite-size systems. PLoS One. 2010;5(12):e14139. https://doi.org/10.1371/journal.pone.0014139 PMID: 21152034
14. Baek SK, Bernhardsson S, Minnhagen P. Zipf's law unzipped. New J Phys. 2011;13(4):043004. https://doi.org/10.1088/1367-2630/13/4/043004
15. Aitchison L, Corradi N, Latham PE. Zipf's law arises naturally when there are underlying, unobserved variables. PLoS Comput Biol. 2016;12(12):e1005110. https://doi.org/10.1371/journal.pcbi.1005110 PMID: 27997544
16. Mazzolini A, Gherardi M, Caselle M, Cosentino Lagomarsino M, Osella M. Statistics of shared components in complex component systems. Phys Rev X. 2018;8(2). https://doi.org/10.1103/physrevx.8.021023
17. Piantadosi ST. Zipf's word frequency law in natural language: a critical review and future directions. Psychon Bull Rev. 2014;21(5):1112–30. https://doi.org/10.3758/s13423-014-0585-6 PMID: 24664880
18. Heaps HS. Information retrieval: computational and theoretical aspects. Academic Press, Inc.; 1978.
19. Cosentino Lagomarsino M, Sellerio AL, Heijning PD, Bassetti B. Universal features in the genome-level evolution of protein domains. Genome Biol. 2009;10(1):R12. https://doi.org/10.1186/gb-2009-10-1-r12 PMID: 19183449
20. Mazzolini A, Colliva A, Caselle M, Osella M. Heaps' law, statistics of shared components, and temporal patterns from a sample-space-reducing process. Phys Rev E. 2018;98(5). https://doi.org/10.1103/physreve.98.052139
21. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36(21):6688–719. https://doi.org/10.1093/nar/gkn668 PMID: 18948295
22. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;5(1):e1000344. https://doi.org/10.1371/journal.pgen.1000344 PMID: 19165319
23. van Nimwegen E. Scaling laws in the functional content of genomes. Trends Genet. 2003;19(9):479–84. https://doi.org/10.1016/S0168-9525(03)00203-8 PMID: 12957540
24. Furusawa C, Kaneko K. Zipf's law in gene expression. Physical Review Letters. 2003;90(8):088102.
25. Koonin EV. Are there laws of genome evolution?. PLoS Computational Biology. 2011;7(8):e1002173. https://doi.org/10.1371/journal.pcbi.1002173
26. Grilli J, Bassetti B, Maslov S, Cosentino Lagomarsino M. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. Nucleic Acids Res. 2012;40(2):530–40. https://doi.org/10.1093/nar/gkr711 PMID: 21937509
27. Lazzardi S, Valle F, Mazzolini A, Scialdone A, Caselle M, Osella M. Emergent statistical laws in single-cell transcriptomic data. Phys Rev E. 2023;107(4–1):044403. https://doi.org/10.1103/PhysRevE.107.044403 PMID: 37198814
28. Grilli J. Macroecological laws describe variation and diversity in microbial communities. Nat Commun. 2020;11(1):4743. https://doi.org/10.1038/s41467-020-18529-y PMID: 32958773
29. Youn H, Bettencourt LMA, Lobo J, Strumsky D, Samaniego H, West GB. Scaling and universality in urban economic diversification. J R Soc Interface. 2016;13(114):20150937. https://doi.org/10.1098/rsif.2015.0937 PMID: 26790997

30. Diestel R. Graph theory. Springer Nature; 2025.
31. Tacchella A, Cristelli M, Caldarelli G, Gabrielli A, Pietronero L. A new metrics for countries' fitness and products' complexity. Sci Rep. 2012;2:723. https://doi.org/10.1038/srep00723 PMID: 23056915
32. Mazzolini A, Caselle M, Osella M. Ranking nodes in bipartite systems with a non-linear iterative map. Commun Phys. 2025;8(1). https://doi.org/10.1038/s42005-025-02073-6
33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;3(Jan):993–1022.
34. Gerlach M, Peixoto TP, Altmann EG. A network approach to topic models. Sci Adv. 2018;4(7):eaaq1360. https://doi.org/10.1126/sciadv.aaq1360 PMID: 30035215
35. Valle F, Osella M, Caselle M. A topic modeling analysis of TCGA breast and lung cancer transcriptomic data. Cancers (Basel). 2020;12(12):3799. https://doi.org/10.3390/cancers12123799 PMID: 33339347
36. Valle F, Caselle M, Osella M. Exploring the latent space of transcriptomic data with topic modeling. NAR Genom Bioinform. 2025;7(2):lqaf049. https://doi.org/10.1093/nargab/lqaf049 PMID: 40264683
37. Valle F, Osella M, Caselle M. Multiomics topic modeling for breast cancer classification. Cancers (Basel). 2022;14(5):1150. https://doi.org/10.3390/cancers14051150 PMID: 35267458
38. Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M. Zipf and Heaps laws from dependency structures in component systems. Phys Rev E. 2018;98(1–1):012315. https://doi.org/10.1103/PhysRevE.98.012315 PMID: 30110773
39. Pang TY, Maslov S. Universal distribution of component frequencies in biological and technological systems. Proc Natl Acad Sci U S A. 2013;110(15):6235–9. https://doi.org/10.1073/pnas.1217795110 PMID: 23530195
40. Altmann EG, Pierrehumbert JB, Motter AE. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. PLoS One. 2009;4(11):e7678. https://doi.org/10.1371/journal.pone.0007678 PMID: 19907645
41. Altmann EG, Cristadoro G, Esposti MD. On the origin of long-range correlations in texts. Proc Natl Acad Sci U S A. 2012;109(29):11582–7. https://doi.org/10.1073/pnas.1117723109 PMID: 22753514
42. Altmann EG, Gerlach M. Statistical laws in linguistics. Lecture Notes in Morphogenesis. Springer; 2016. p. 7–26. https://doi.org/10.1007/978-3-319-24403-7_2
43. Brown JH. Macroecology. University of Chicago Press; 1995.
44. Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005;46(5):323–51. https://doi.org/10.1080/00107510500052444
45. Mazzarisi O, de Azevedo-Lopes A, Arenzon JJ, Corberi F. Maximal Diversity and Zipf's Law. Phys Rev Lett. 2021;127(12):128301. https://doi.org/10.1103/PhysRevLett.127.128301 PMID: 34597111
46. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. Internet Mathematics. 2004;1(2):226–51. https://doi.org/10.1080/15427951.2004.10129088
47. Herdan G. Quantitative Linguistics. Butterworths; 1964.
48. Egghe L. Untangling Herdan's law and Heaps' law: mathematical and informetric arguments. J Am Soc Inf Sci. 2007;58(5):702–9. https://doi.org/10.1002/asi.20524
49. Tria F, Loreto V, Servedio VDP, Strogatz SH. The dynamics of correlated novelties. Sci Rep. 2014;4:5890. https://doi.org/10.1038/srep05890 PMID: 25080941
50. Loreto V, Servedio VD, Strogatz SH, Tria F. Dynamics on expanding spaces: modeling the emergence of novelties. Creativity and universality in language. 2016. p. 59–83.
51. Yang VC, Holehouse J, Youn H, Arroyo JI, Redner S, West GB, et al. Scaling laws for function diversity and specialization across socioeconomic and biological complex systems. Proc Natl Acad Sci U S A. 2026;123(7):e2509729123. https://doi.org/10.1073/pnas.2509729123 PMID: 41678302
52. Gerlach M, Altmann EG. Scaling laws and fluctuations in the statistics of word frequencies. New J Phys. 2014;16(11):113010. https://doi.org/10.1088/1367-2630/16/11/113010
53. Tria F, Crimaldi I, Aletti G, Servedio VDP. Taylor's Law in Innovation Processes. Entropy (Basel). 2020;22(5):573. https://doi.org/10.3390/e22050573 PMID: 33286342
54. Taylor LR. Aggregation, Variance and the Mean. Nature. 1961;189(4766):732–5. https://doi.org/10.1038/189732a0
55. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome Biol Evol. 2013;5(1):233–42. https://doi.org/10.1093/gbe/evt002 PMID: 23315380
56. Elhanati Y, Sethna Z, Callan Jr CG, Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev. 2018;284(1):167–79. https://doi.org/10.1111/imr.12665 PMID: 29944757
57. Ruiz Ortega M, Spisak N, Mora T, Walczak AM. Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals. PLoS Genet. 2023;19(2):e1010652. https://doi.org/10.1371/journal.pgen.1010652 PMID: 36827454
58. Grilli J, Romano M, Bassetti F, Cosentino Lagomarsino M. Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers. Nucleic Acids Res. 2014;42(11):6850–60. https://doi.org/10.1093/nar/gku378 PMID: 24829449

59. Gerlach M, Font-Clos F, Altmann EG. Similarity of symbol frequency distributions with heavy tails. Phys Rev X. 2016;6(2). https://doi.org/10.1103/physrevx.6.021009
60. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Dewhurst DR, et al. Allotaxonometry and rank-turbulence divergence: a universal instrument for comparing complex systems. EPJ Data Sci. 2023;12(1). https://doi.org/10.1140/epjds/s13688-023-00400-x
61. Blumm N, Ghoshal G, Forró Z, Schich M, Bianconi G, Bouchaud J-P, et al. Dynamics of ranking processes in complex systems. Phys Rev Lett. 2012;109(12):128701. https://doi.org/10.1103/PhysRevLett.109.128701 PMID: 23005999
62. Iñiguez G, Pineda C, Gershenson C, Barabási A-L. Dynamics of ranking. Nat Commun. 2022;13(1):1646. https://doi.org/10.1038/s41467-022-29256-x PMID: 35347126
63. Lenz M, Müller F-J, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci Rep. 2016;6:25696. https://doi.org/10.1038/srep25696 PMID: 27254731
64. Heimberg G, Bhatnagar R, El-Samad H, Thomson M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. Cell Syst. 2016;2(4):239–50. https://doi.org/10.1016/j.cels.2016.04.001 PMID: 27135536
65. Biondo M, Cirone N, Valle F, Lazzardi S, Caselle M, Osella M. The intrinsic dimension of gene expression during cell differentiation. Nucleic Acids Res. 2025;53(16):gkaf805. https://doi.org/10.1093/nar/gkaf805 PMID: 40889499
66. Staiano M, Cirone N, Biondo M, Osella M, Scialdone A. Intrinsic dimensionality of single-cell transcriptomic data reveals potency landscapes during cell reprogramming. bioRxiv. 2025;:2025–07. https://doi.org/10.1101/2025.07.21.665922
67. Rouzine IM, Rozhnova G. Antigenic evolution of viruses in host populations. PLoS Pathog. 2018;14(9):e1007291. https://doi.org/10.1371/journal.ppat.1007291 PMID: 30208108
68. Chardès V, Mazzolini A, Mora T, Walczak AM. Evolutionary stability of antigenically escaping viruses. Proc Natl Acad Sci U S A. 2023;120(44):e2307712120. https://doi.org/10.1073/pnas.2307712120 PMID: 37871216
69. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004;305(5682):371–6. https://doi.org/10.1126/science.1097211 PMID: 15218094
70. Fonville JM, Wilks SH, James SL, Fox A, Ventresca M, Aban M, et al. Antibody landscapes after influenza virus infection or vaccination. Science. 2014;346(6212):996–1000. https://doi.org/10.1126/science.1256427 PMID: 25414313
71. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. Integrating influenza antigenic dynamics with molecular evolution. Elife. 2014;3:e01914. https://doi.org/10.7554/eLife.01914 PMID: 24497547
72. Serrano MA, Flammini A, Menczer F. Modeling statistical properties of written text. PLoS One. 2009;4(4):e5372. https://doi.org/10.1371/journal.pone.0005372 PMID: 19401762
73. Ebeling W, Pöschel T. Entropy and long range correlations in literary English. Europhys Lett. 1994;26(2):241.
74. Church KW, Gale WA. Poisson mixtures. Nat Lang Eng. 1995;1(2):163–90. https://doi.org/10.1017/s1351324900000139
75. Kleinberg J. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002. 91–101.
76. Sireci M, Muñoz MA, Grilli J. Environmental fluctuations explain the universal decay of species-abundance correlations with phylogenetic distance. Proc Natl Acad Sci U S A. 2023;120(37):e2217144120. https://doi.org/10.1073/pnas.2217144120 PMID: 37669363
77. Molina N, van Nimwegen E. Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet. 2009;25(6):243–7. https://doi.org/10.1016/j.tig.2009.04.004 PMID: 19457568
78. De Lazzari E, Grilli J, Maslov S, Cosentino Lagomarsino M. Family-specific scaling laws in bacterial genomes. Nucleic Acids Res. 2017;45(13):7615–22. https://doi.org/10.1093/nar/gkx510 PMID: 28605556
79. Maslov S, Krishna S, Pang TY, Sneppen K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc Natl Acad Sci U S A. 2009;106(24):9743–8. https://doi.org/10.1073/pnas.0903206106 PMID: 19482938
80. Molina N, van Nimwegen E. The evolution of domain-content in bacterial genomes. Biol Direct. 2008;3:51. https://doi.org/10.1186/1745-6150-3-51 PMID: 19077245
81. Park J, Newman MEJ. Statistical mechanics of networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;70(6 Pt 2):066117. https://doi.org/10.1103/PhysRevE.70.066117 PMID: 15697444
82. Cimini G, Squartini T, Saracco F, Garlaschelli D, Gabrielli A, Caldarelli G. The statistical physics of real-world networks. Nat Rev Phys. 2019;1(1):58–71. https://doi.org/10.1038/s42254-018-0002-6
83. van Leijenhorst DC, Van der Weide TP. A formal derivation of Heaps' Law. Information Sciences. 2005;170(2–4):263–72.
84. Eliazar I. The growth statistics of Zipfian ensembles: Beyond Heaps' law. Physica A: Statistical Mechanics and its Applications. 2011;390(20):3189–203. https://doi.org/10.1016/j.physa.2011.05.003
85. Bianconi G, Pin P, Marsili M. Assessing the relevance of node features for network structure. Proc Natl Acad Sci U S A. 2009;106(28):11433–8. https://doi.org/10.1073/pnas.0811511106 PMID: 19571013
86. Leinaas JM, Myrheim J. On the theory of identical particles. Il Nuovo Cimento B. 1977;37(1):1–23.
87. Bianconi G, Barabási AL. Bose-Einstein condensation in complex networks. Phys Rev Lett. 2001;86(24):5632–5. https://doi.org/10.1103/PhysRevLett.86.5632 PMID: 11415319

88. Bassetti B, Zarei M, Cosentino Lagomarsino M, Bianconi G. Statistical mechanics of the "Chinese restaurant process": lack of self-averaging, anomalous finite-size effects, and condensation. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;80(6 Pt 2):066118. https://doi.org/10.1103/PhysRevE.80.066118 PMID: 20365242
89. Ferrer R, Cancho R. The variation of Zipf's law in human language. Eur Phys J B. 2005;44:249–57.
90. De Marzo G, Gabrielli A, Zaccaria A, Pietronero L. Quantifying the unexpected: a scientific approach to Black Swans. Phys Rev Research. 2022;4(3). https://doi.org/10.1103/physrevresearch.4.033079
91. Purves DW, Pacala SW. Ecological drift in niche-structured communities: neutral pattern does. Biotic interactions in the tropics: their role in the maintenance of species diversity. 2005. p. 107.
92. Haegeman B, Weitz JS. A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics. 2012;13:196. https://doi.org/10.1186/1471-2164-13-196 PMID: 22613814
93. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol. 2001;313(4):673–81. https://doi.org/10.1006/jmbi.2001.5079 PMID: 11697896
94. Quake SR. The cell as a bag of RNA. Trends Genet. 2021;37(12):1064–8. https://doi.org/10.1016/j.tig.2021.08.003 PMID: 34462156
95. Corominas-Murtra B, Hanel R, Thurner S. Understanding scaling through history-dependent processes with collapsing sample space. Proc Natl Acad Sci U S A. 2015;112(17):5348–53. https://doi.org/10.1073/pnas.1420946112 PMID: 25870294
96. Iacopini I, Di Bona G, Ubaldi E, Loreto V, Latora V. Interacting discovery processes on complex networks. Phys Rev Lett. 2020;125(24):248301. https://doi.org/10.1103/PhysRevLett.125.248301 PMID: 33412072
97. Schwab DJ, Nemenman I, Mehta P. Zipf's law and criticality in multivariate data without fine-tuning. Phys Rev Lett. 2014;113(6):068102. https://doi.org/10.1103/PhysRevLett.113.068102 PMID: 25148352
98. Marsili M, Mastromatteo I, Roudi Y. On sampling and modeling complex systems. J Stat Mech. 2013;2013(09):P09003. https://doi.org/10.1088/1742-5468/2013/09/p09003
99. Marsili M, Roudi Y. Quantifying relevance in learning and inference. Physics Reports. 2022;963:1–43. https://doi.org/10.1016/j.physrep.2022.03.001
100. Mikhaylovskiy N. Zipf's and Heaps' Laws for Tokens and LLM-generated Texts. In: Findings of the Association for Computational Linguistics: EMNLP 2025, 2025. p. 15469–81. https://doi.org/10.18653/v1/2025.findings-emnlp.837
101. Reviriego P, Conde J, Merino-Gómez E, Martínez G, Hernández JA. Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. Machine Learning with Applications. 2024;18:100602. https://doi.org/10.1016/j.mlwa.2024.100602
102. Lupyan G. The unreasonable effectiveness of pattern matching. arXiv preprint. 2026. https://arxiv.org/abs/2601.11432
103. Power A, Burda Y, Edwards H, Babuschkin I, Misra V. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv preprint. 2022. https://arxiv.org/abs/2201.02177
104. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint. 2022. https://arxiv.org/abs/2206.07682