OXFORD

Integrative Biology, 2021, **13**, 283–294
https://doi.org/10.1093/intbio/zyab020
Advance access publication date 22 December 2021
Original Article

Evolution and emergence: higher order information structure in protein interactomes across the tree of life

Brennan Klein[1,2,*], Erik Hoel[3], Anshuman Swain[4], Ross Griebenow[5] and Michael Levin[3]

[1]Network Science Institute, Northeastern University, Boston, MA, USA
[2]Laboratory for the Modeling of Biological and Socio-Technical Systems, Northeastern University, Boston, MA, USA
[3]Allen Discovery Center, Tufts University, Medford, MA, USA
[4]Department of Biology, University of Maryland, College Park, MD, USA
[5]Department of Computer Science, Drexel University, Philadelphia, PA, USA
*__Corresponding author__. E-mail: b.klein@northeastern.edu

Abstract

The internal workings of biological systems are notoriously difficult to understand. Due to the prevalence of noise and degeneracy in evolved systems, in many cases the workings of everything from gene regulatory networks to protein–protein interactome networks remain black boxes. One consequence of this black-box nature is that it is unclear at which scale to analyze biological systems to best understand their function. We analyzed the protein interactomes of over 1800 species, containing in total 8 782 166 protein–protein interactions, at different scales. We show the emergence of higher order 'macroscales' in these interactomes and that these biological macroscales are associated with lower noise and degeneracy and therefore lower uncertainty. Moreover, the nodes in the interactomes that make up the macroscale are more resilient compared with nodes that do not participate in the macroscale. These effects are more pronounced in interactomes of eukaryota, as compared with prokaryota; these results hold even after sensitivity tests where we recalculate the emergent macroscales under network simulations where we add different edge weights to the interactomes. This points to plausible evolutionary adaptation for macroscales: biological networks evolve informative macroscales to gain benefits of both being uncertain at lower scales to boost their resilience, and also being 'certain' at higher scales to increase their effectiveness at information transmission. Our work explains some of the difficulty in understanding the workings of biological networks, since they are often most informative at a hidden higher scale, and demonstrates the tools to make these informative higher scales explicit.

Keywords: emergence, evolution, interactome, network

Insight Box

Protein interactomes are an insight into the inner workings of cells. Here we analyze the protein interactomes of over 1800 different species using information theory to measure the amount of noise and uncertainty in protein interactions. We found that uncertainty associated with protein interactions is higher in eukaryotes compared with prokaryotes. To explore why this is the case, we also modeled the same protein interactomes at higher scales by coarse graining. We found that eukaryotes had much more informative higher scales in their interactomes (in the form of less noise and uncertainty). We explore the benefits of such higher scale structure and examine the relationship between evolution, information and scale in biological systems.

INTRODUCTION

Interactions in biological systems are noisy and degenerate in their functions, making them fundamentally noisier and fundamentally different from those in engineered systems [1, 2]. The sources of noise in biology are nearly ubiquitous and vary widely. Noise may exist a gene regulatory network, wherein a gene might upregulate another gene but only probabilistically, or they may be noisy in that a protein may bind randomly across a set of possible pairings. There are numerous sources of such indeterminism in cells and tissues, such as how cell molecules are buffered by Brownian motion [3], to the stochastic opening and closing of ion channels [4], and even to the chaotic dynamics of neural activity [5].

There are also numerous sources of degeneracy within the cellular, developmental and genetic operation of organisms [6]. Degeneracy is when an end state or output, like a phenotype, can come from a large number of possible states or inputs [7].

Due to this indeterminism and degeneracy, the dynamics and function of biological systems are often uncertain. This hampers control of system-level

Received: June 25, 2021. **Revised:** November 16, 2021. **Editorial decision:** November 20, 2021. **Accepted:** November 25, 2021

properties for biomedicine and synthetic bioengineering, as well as hampering the understanding of modelers and experimentalists who wish to build 'big data' approaches to biology like interactomes, connectomes and mapping molecular pathways [8, 9]. Although there have been many attempts to characterize and understand this uncertainty in biological systems [7], the explanations typically do not extend beyond the advantages of redundancy in these systems [10].

How do noise and uncertainty span the tree of life? Here we examine this question in biological networks, a common type of model for biological systems [11, 12]. Specifically, we examine protein–protein interactomes from organisms across a wide range of organisms to investigate whether or not the noise and uncertainty in biological networks increases or decreases across evolution. To quantify this noise and uncertainty, we make use of the *effective information* (*EI*), an information-theoretic network quantity based on the entropy of random walker behavior on a network. A lower *EI* indicates greater noise and uncertainty (a formal mathematical definition is given in the section 'Results'). Indeed, the *EI* of biological networks has already been shown to be lower in biological networks compared with technological networks [13], which opens the question of why this is the case.

To see how *EI* changes across evolution, we examined networks of protein–protein interactions (PPIs) from organisms across the tree of life. The dataset consists of interactomes from 1840 species (1539 bacteria, 111 archaea and 190 eukaryota) derived from the STRING database [14, 15]. These interactomes have been previously used to study evolution of resilience, where researchers found that species tended to have higher values of network resilience with increasing evolution (wherein 'evolution' was defined as the number of nucleotide substitutions per site) [16]. In our work, we take a similar approach, highlighting changes in interactome properties as evolution progresses.

In addition, we focus on identifying when interactomes have informative macroscales. A *macroscale* refers to some dimension reduction, such as an aggregation, coarse graining or grouping, of states or elements of the biological system. In networks, this takes the form of replacing subgraphs of the network with individual nodes (macro-nodes). A network has an informative macroscale when subgraphs of the network can be grouped into macro-nodes such that the resulting dimensionally reduced network gains *EI* [13]. When such grouping leads to an increase in EI, we describe the resulting macro-node as being part of an informative macroscale. Following previous work, we refer to any gain in *EI* at the macroscale as causal emergence [17]. With these techniques, we can identify which PPI networks have informative macroscales and which do not. By correlating this property with where (in time) each species lies in the evolutionary tree, we show that informative macroscales tend to emerge later in evolution, being associated more with eukaryota than prokaryota (such as bacteria). Using a number of sensitivity and robustness tests, we show that these results are not explained by other network properties such as network density or size.

What is the evolutionary advantage of having informative higher scales? This question is important because higher scales minimize noise or uncertainty in biological networks. Yet such uncertainty or noise represents a fundamental paradox. The more noisy a network is, the more uncertain and the less effective that network is; effectiveness, here, refers to the ability to reliably transform inputs into outputs (similar to the notion of specificity), such as being able to upregulate a particular gene in response to a detected chemical in the environment. In this sense, we might expect evolved networks to be highly effective—that is, we might expect them to have structures that reliably produce specific outputs given a certain set of inputs/causes. Yet this is the opposite of what we observe. Instead, we observe that effectiveness of lower scales decreases later in evolution, as higher scales that are effective emerge.

We argue here that this multiscale behavior is the resolution to a paradox: there are advantages to being effective, but there are also advantages to being less effective and therefore more uncertain or noisy. For instance, less effective networks might be more resistant to attack or node failure due to redundancy. The paradox is that networks that are certain are effective yet are vulnerable to attacks or node failures, whereas networks that are uncertain are less effective but are resilient in the face of attacks or node failures. We argue that biological networks have evolved to resolve this 'certainty paradox' by having informative higher scales. Specifically, we propose that the macroscales of a biological network evolve to have high effectiveness, but their underlying microscales may have low effectiveness, therefore making the system resilient without paying the price of a low effectiveness.

In a biological sense, node failures or attacks in a cellular network may represent certain mutations in proteins or other biochemical entities, which in turn may prevent regular functioning of the system [18]. Biological networks should then, over the course of evolution, develop degeneracy and noise at lower scales to maintain regular functioning, while at the same time developing effectiveness at a higher level. This transformation can be achieved by the action of both neutral and selective processes in evolution. Neutral processes such as presuppression, which, aided by mutations, increases the number of interactions [19] and can therefore decrease network effectiveness. On the other hand, selective processes can weed out the noise that interferes with the functioning and efficiency of the system [20]. An interplay of these evolutionary processes can lead to a resolution of the 'certainty paradox' in cellular networks by the development of informative macroscales.

This work therefore presents an explanation for the observed trend in increased resiliency through evolution [16]: informative macroscales make networks more resilient. Finally, we offer insights into biological processes at molecular level that might be responsible for the emergence of informative macroscales in PPI networks, specifically looking at the differences between bacteria, which have a low rate of nucleotide substitutions per site, and eukaryota, which exhibit a higher rate. Understanding the basic principles governing the differences in efficiency and uncertainty between these major divisions of life can help us comprehend the trade-offs involved in information processes in PPIs across evolution.

RESULTS

Effectiveness of protein interactomes across the tree of life

EI is a network property reflecting the certainty (or uncertainty) contained in that network's connectivity [13]. It is a structural property of a network calculated by traversing its topology and is based on the uncertainty of a random walker's transitions between pairs of nodes and the distribution of this uncertainty throughout the network. It is calculated by examining the network's connectivity.

In a protein interactome, the nodes are individual proteins, and the edges of the network are interactions, generally describing the possibility of binding between two proteins. Therefore, the uncertainty we analyze is uncertainty as to which protein(s) a given protein might interact with, e.g. bind with. Each node in the network has out-weights, which are represented by a vector, W_i^{out}. For instance, protein *A* might share an edge with protein *B* and also protein *C*. Therefore, W_A^{out} is [½, ½]. Since most protein interactomes are undirected, its edges are normalized for each node (such that the sum of W_i^{out} for each node is 1.0). Note that this process of normalization implies that the probability of binding is uniform across the different possible interactions. This transformation into a direct network makes the networks amenable to standard tools of network science, such as analyzing random walk dynamics, and it is also necessary to calculate the *EI* of the network. In addition, the uniform distribution of $^1/_n$ is the simplest a priori assumption. However, the actual probability of binding is dependent on biological background conditions such as protein prevalence and not included in most open-source models, and therefore our analysis could change if such detailed probabilities were known.

The uncertainty associated with each protein can be mathematically captured by examining the entropy of the outputs of a node, $H(W_i^{out})$, wherein a higher entropy indicates more uncertainty as to interactions [21]. The entropy of the distribution of weight across the entire network, $H(\langle W_i^{out}\rangle)$, reflects the spread of uncertainty across the network. A lower $H(\langle W_i^{out}\rangle)$ means that information is distributed only over a small number of nodes. A high $H(\langle W_i^{out}\rangle)$ signifies that information is dispersed throughout the network. The *EI* of a network can then be defined as the entropy of distribution of weights over the network minus the average uncertainty inherent in the weight of each node, or:

$$\mathrm{EI} = \mathrm{H}\left(\langle W_i^{out}\rangle\right) - \langle H\left(W_i^{out}\right)\rangle \qquad (1)$$

EI can itself be further decomposed into the degeneracy and indeterminism of a network [13, 22], where each indicates the lack of specificity in the network's connectivity or interactions. Degeneracy indicates a lack of specificity in targeting nodes (many nodes target the same node), whereas indeterminism indicates a lack of specificity in targeted nodes (nodes target many nodes). In network science, a node that is connected to many other nodes is said to have a high degree, whereas here in this context we refer to it as having high indeterminism; both terms refer to the same structural property, but in this context, we view nodes with high degree as sources of uncertainty in the network (i.e. a random walker on a high-degree node is more uncertain about which node it will visit next compared with a random walker on a low-degree node). Note that, if networks are considered deterministic in the physical sense, the indeterminism term of *EI* still reflects the uniqueness of targets in the network.

A network where all the nodes target a single node will have zero *EI* (since it has maximum degeneracy), as will a network where all nodes target all other nodes (complete indeterminism). *EI* will only be maximal if every node has a unique output. This forces the *EI* of a network to be bounded by $\log_2(n)$, where n is the number of nodes in the network. Therefore, to compare networks of different sizes, *EI* can be normalized, and a new quantity, effectiveness, can be defined as:

$$\mathbf{effectiveness} = EI/\mathbf{log}_2(n) \qquad (2)$$

Again, we use the term 'effectiveness' in a way that is slightly different from its colloquial meaning; we instead use it to quantify the certainty that a given set of outputs will follow a particular set of inputs. To explore the change in effectiveness of biological networks, we examined protein interactomes of 1840 species divided between archaea, bacteria and eukaryota (see section 'Methods' for details on the origin and nature of these protein interactomes). We found a clear pattern in the effectiveness of the networks, based on where they are located in the tree of life (Fig. 1), the position of which is based on each protein interactome's small subunit ribosomal RNA gene sequence information [23] (see section 'Methods' for details). Overall, we found that the mean effectiveness of protein interactomes decreases later in the tree of life as nucleotide substitutions occurred. Specifically, bacteria were found to have a greater effectiveness (0.77) compared with eukaryota (0.72) on average (Student's *t*-test, $P < 10^{-8}$). Following Zitnik *et al.*, we restricted further statistical analysis to interactomes with >1000 citations, to use the most well-founded protein interactomes, but the directionality

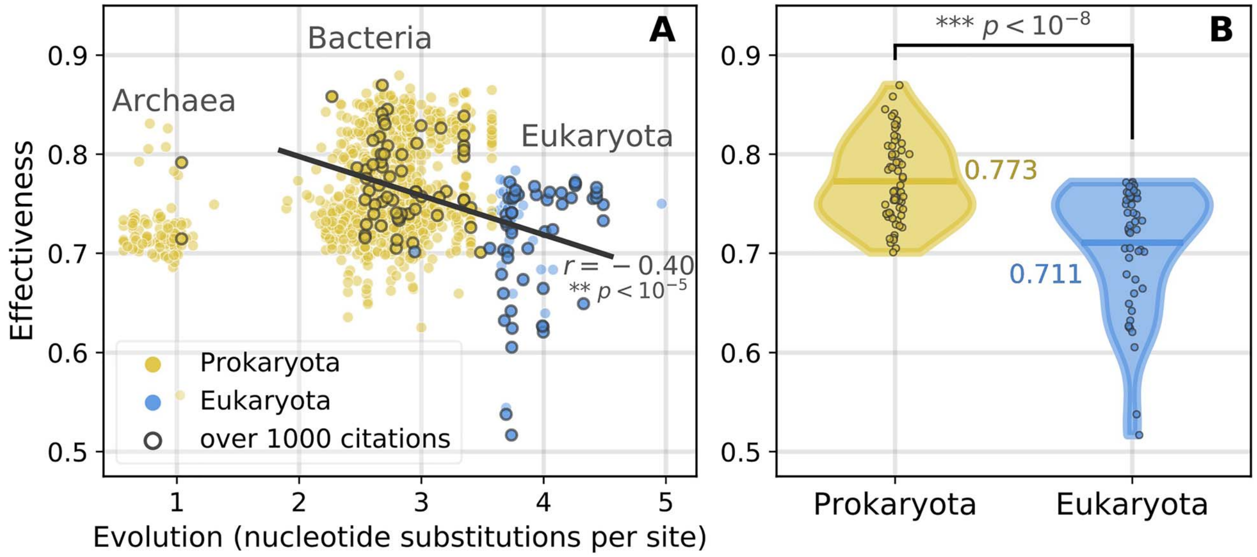

Figure 1. Effectiveness of protein interactomes. (A) Effectiveness of all 1840 species with their superphylum association. Interactomes with a lower number of nucleotide substitutions per site tended to be prokaryota (yellow), whereas those higher tended to be eukaryota (blue). Solid line is a linear regression comparing the effectiveness of bacteria and eukaryota ($r = -0.40$, $P < 10^{-5}$), due to the small number of archaea that passed the threshold for reliable datasets (see section 'Results'). (B) The effectiveness of prokaryotic protein interactomes is greater than that of eukaryotic species, indicating that effectiveness might decrease with more nucleotide substitutions per site.

and significance of the result are unchanged when all interactomes are included (Student's *t*-test, $P < 10^{-11}$), indicating that the selection procedure of only using the most-cited organisms does not influence our results. Due to the small number of archaea interactomes based on above 1000 citations, we did not include those samples in Figure 1B.

Causal emergence across the tree of life

At first the higher effectiveness in prokaryota interactomes as compared to that of eukaryota (as shown in Fig. 1) may seem counter intuitive. One might naively expect the effectiveness of cellular machinery, including or especially interactomes, to increase over evolutionary time, instead of decreasing as we have shown.

One hypothesis to explain these results is that, although the protein interactomes get less effective in their microscales over evolutionary time, the interactomes are able to nonetheless be effective due to the emergence of informative macroscales as evolution proceeds. To examine this hypothesis, we must first define a procedure for finding macroscales in networks.

Network macroscales are defined as subgraphs (i.e. connected sets of nodes and their associated links) that can be grouped into single macro-nodes such that the resulting network has a higher value of *EI* than the original microscale network [13]. We denote the microscale of a network as G and the macroscale as G_M, which is composed of both ungrouped nodes (micro-nodes) and macro-nodes, μ. The macroscale network, G_M, is a dimension reduction in that it always has fewer nodes than G.

To recast a particular subgraph into a macro-node, its connectivity must be modified since the subgraph is being transformed into a single node. In terms of input to the new macro-node, μ, all out-weights that targeted nodes in the subgraph now target the macro-node. In terms of output, each micro-node, v_i inside the subgraph has some W_i^{out}. To recast the nodes inside a subgraph into a macro-node, we replace the W_i^{out} of the nodes in the subgraph with a single W_μ^{out}, which is a weighted average of the set of each W_i^{out} in the subgraph. The weight is based on the probability P of each node v_i in the stationary distribution of the network, π. This forms macro-nodes $(\mu|\pi)$ that accurately recapitulate the microscale random walk dynamics at the new macroscale [13]. A macroscale is informative if it increases the *EI* of the network compared with the original microscale. To find macro-nodes that maximally increase *EI*, we make use of a modified spectral algorithm to find locally optimal micro-to-macro mappings, originally described in Griebenow *et al.* [24].

Results from this analysis support our initial hypothesis that effectiveness is actually being transitioned to macroscales of biological networks in eukaryota over evolutionary time, even though the microscales become noisy and less effective over evolutionary time. The total amount of causal emergence (the gain of *EI* by grouping subgraphs into macro-nodes) was identified for each protein interactome from each species, normalized by the total size of that protein interactome (Fig. 2B). Across the tree of life, we observe that eukaryota have more informative macroscales and show a significant difference in the percentage of microscale nodes that get grouped into macro-nodes than prokaryota (Fig. 2A).

Macroscales of interactomes are more resilient than microscales

Ultimately, although there may be many reasons to see an evolutionary increase in causal emergence, here, we explore the evidence for a specific benefit of having

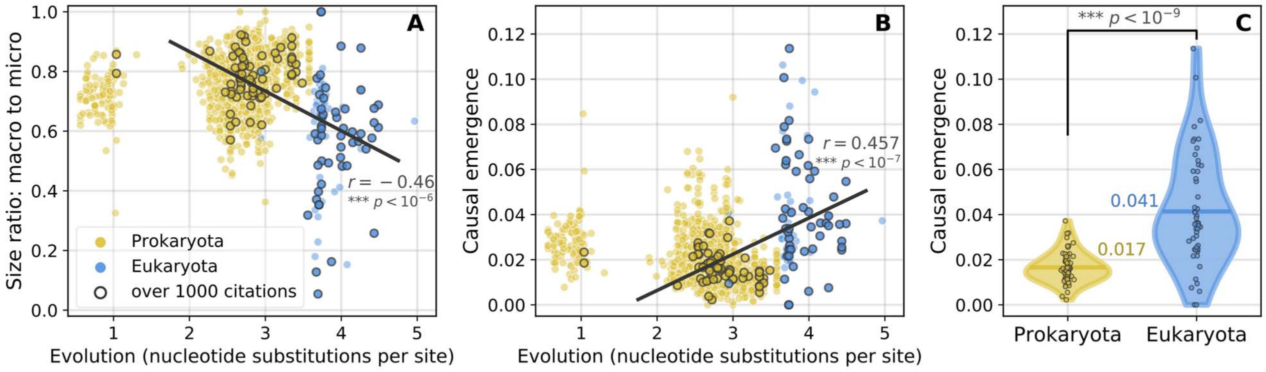

Figure 2. Causal emergence in protein interactomes. (A) The protein interactomes of each species undergoes a modified spectral analysis to identify the scale with $EI^{\max}$. The total dimension reduction of the network is shown, with there being a greater effect in eukaryota as more subgraphs are grouped in macro-nodes. That is, as evolutionary time goes on the coarse-grained networks become a smaller fraction of their original microscale network size ($r = -0.46$, $P < 10^{-6}$). (B) To compare the degree of causal emergence in protein interactomes of different sizes, the total amount of causal emergence is normalized by the size of the network, $\log_2(n)$, and we see here a positive correlation between evolution and causal emergence ($r = 0.457$, $P < 10^{-7}$), results which do not change in directionality or significance when using either full data set of organisms or just the top 100 most cited. (C) The amount of normalized causal emergence is significantly higher for eukaryota.

multiscale structure, which networks with only a single scale lack. This specific benefit comes from the fact that all networks face a 'certainty paradox.' The paradox is that uncertainty in connectivity is desirable since it is protective from node failures. For instance, a node failure could be the removal of a protein due to a nonsense mutation, or the inability to express a certain protein due to an environmental effect, such as a lack of resources, or even a viral attack. In turn, this could lead to a loss of biological function or the development of disease or even cell death. A protein interactome may be resilient to such node failures by being highly uncertain or degenerate in its protein–protein interactomes. However, this comes at a cost. A high uncertainty can lead to problems with reliability, uniqueness and control in terms of effects, such as an inability for a particular protein to deterministically bind with another protein. For instance, in a time of environmental restriction of resources, certain protein–protein interactomes may be necessary for continued cellular function, but if there is large-scale uncertainty even significant upregulation of genes controlling expression may not lead reliably to a certain interaction.

Here we explore these issues by examining the network resilience of protein interactomes in response to node removals, which represent either attacks or general node failures. To measure the resilience of the network in response to a node removal we follow [16] by using the change in the Shannon entropy of the component size distribution of the network following random node removal. That is, if p_c is the probability that a randomly selected node is in connected component $c \in C$ following the removal of a fraction f of the nodes in the network, the entropy associated with the component size distribution, $H(G_f)$, is:

$$H\left(G_f\right) = -\frac{1}{\log_2(N)} \sum_{c}^{n_c} p_c \log_2\left(p_c\right) \tag{3}$$

where n_c is the number of connected components remaining after f fraction nodes have been removed (note: 'removed' here indicates that the nodes become isolates, still contributing to the component size distribution though not retaining any of the original links). The change in entropy, $H(G_f)$, as f from 0.0 to 1.0 corresponds to the resilience of the network in question. Specifically, this resilience is defined as follows:

$$\text{Resilience}(G) = 1 - \sum_{f=0}^{1} \frac{H\left(G_f\right)}{r_f} \tag{4}$$

where r_f is the rate of node removal (i.e. the increment that the fraction f increases from 0 to 1). In this work, we default to a value of $r_f = 100$, which means that the calculation of a network's resilience involves iteratively removing 1, 2, . . . 100% of the nodes in the network. For each value of f, we simulate the node removal process 20 times.

Our hypothesis is that biological networks deal with this 'certainty paradox' by maintaining uncertainty at their microscale. This gives a pool of noise and degeneracy, leading to resilience. Meanwhile, at the macroscale, the networks can develop a high effectiveness, wherein sets of proteins deterministically and non-degenerately interact. To explore this hypothesis, we compare the network's resilience to removing micro-nodes that are members of subgraphs grouped into macro-nodes to the network's resilience to removing micro-nodes that remain ungrouped (shown in Fig. 3).

By isolating the calculation of network resilience to only the micro- or macro-nodes of a network, we see a stark trend emerge wherein nodes inside highly informative macro-nodes are more resilient than nodes outside. That is, nodes in the original interactome that were grouped into a macro-node contribute more to the overall resilience of the interactome. This not only supports our hypothesis that biological networks resolve the 'certainty

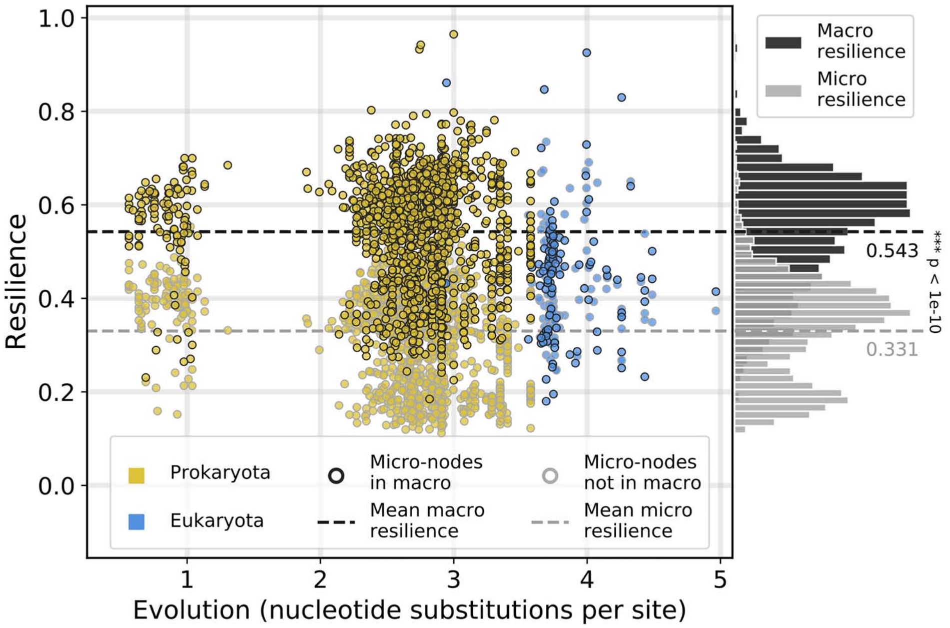

Figure 3. Resilience of micro- and macro-nodes following causal emergence in interactomes. The resilience of a species interactome changes across the tree of life, as shown in previous research [16]. Using the mapping generated by computing causal emergence (Fig. 2B), we calculate the resilience of the network, isolating the calculation to nodes that are either part of the macroscale or microscale. Points are color-coded according to the evolutionary domain; points with dark outlines are associated with micro-nodes that have been grouped into a macro-node (macroscale), whereas the points with light outlines have not been grouped into a macro-node (microscale). Nodes at the microscale contribute less to the overall resilience of a given network (0.331) compared with nodes that contribute to macro-nodes (0.543) on average (*t*-test, $P < 10^{-10}$). Note: plotted are the microscale and macroscale resilience values for each interactome in the dataset, and the difference in resilience across scales holds even when only including species with >10, 100 or 1000 citations.

paradox' by building multiscale structure, but also provides further explanation and contextualization for the recent findings of increasing resilience across evolutionary time [16].

DISCUSSION

In this work, we analyzed how the informativeness of protein interactomes changed over evolutionary time. Specifically, we made use of the *EI* to analyze the amount of uncertainty (or noise) in the connectivity of protein interactomes. We found that the effectiveness (the normalized *EI*) of protein interactomes decreased over evolutionary time, indicating that uncertainty in the connectivity of the interactomes was increasing over evolutionary time. However, we discovered that this was due to eukaryotic protein interactomes possessing higher (informative) scales, such that they had more *EI* when recast as a coarse-grained network—a phenomenon known as causal emergence. This lower effectiveness and higher causal emergence in eukaryotic species was due to the indeterminism and degeneracy in the network structure of their PPIs.

We used a dataset from the STRING database [14, 15] that spans >1800 species (1539 bacteria, 111 archaea and 190 eukaryota), which has been shown to have considerable advantages compared with previous collections of protein interactomes [16]. However, we cannot rule out the possibility that biases might exist in the specific manner of data collection, such as high under representation of specific types of difficult-to-detect interactions, which could potentially introduce errors in the calculations of effectiveness in eukaryotic interactomes. As such, we conducted a series of statistical robustness tests that accounted for potential biases in both the data collection and network structures of interactomes in our dataset (see Fig. 4 in section 'Methods' for further details about these statistical tests). In short, the results we observed in this study cannot be explained by two plausible sources of bias: (1) Random rewiring of network edges does not produce similar results and (2) Network null models of each interactome in this study produce only a fraction of the observed causal emergence in our dataset (the maximum causal emergence values for a species' network null model only reached 3% of the causal emergence of the original interactome). Notwithstanding these statistical tests, as technology and methods continue to improve, these results and hypotheses should continue to be tested.

Macroscales themselves may be important both from the perspective of the studying biologist understanding what the intrinsic or functionally relevant scales of a biological system are and also from the perspective of the system themselves. For instance, possessing informative macroscales might make for easier control of particular outcomes or processes, such as cell differentiation; they may also be easier for evolution to construct or evolve, since macroscales are by definition multiply realizable.

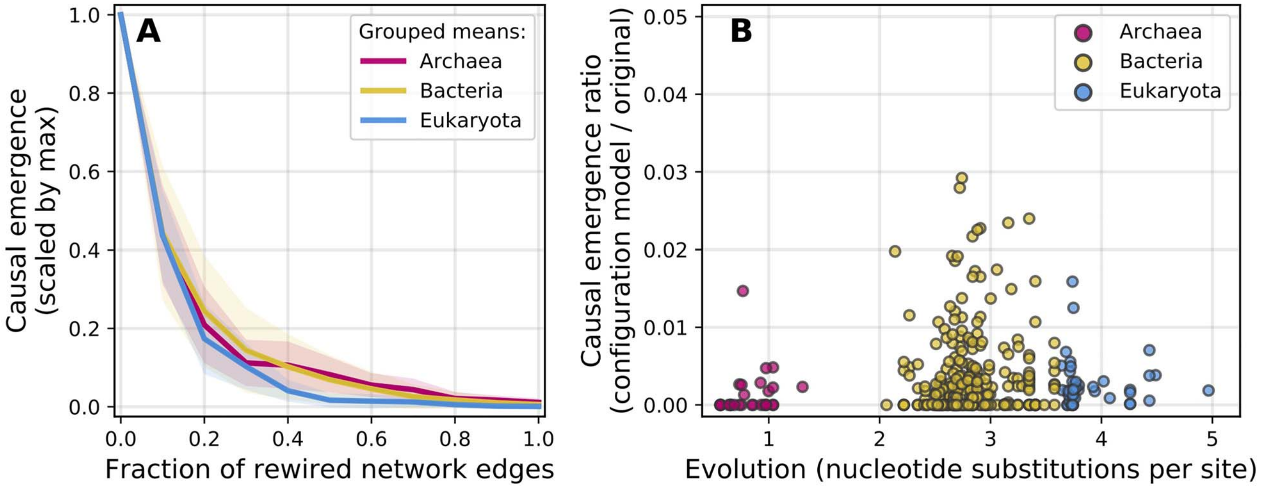

Figure 4. Statistical controls and network robustness tests. (A) As a greater fraction of network links are randomly rewired, we observe a decrease in the causal emergence values of the resulting networks (normalized by the causal emergence value of the original network). This appears not to be dependent on evolutionary domain, network size, density or other network properties. Error bands are 95% confidence intervals. (B) A second statistical control known as a soft configuration model assesses whether there is anything intrinsic to the network's degree distribution that could be driving a given result. Here, we divide the average causal emergence of 10 such configuration model networks by the causal emergence values of the original protein interactome and observe that the null model networks preserve only a small fraction of the original amount of information gain (at most, the configuration models may show 3% of the original causal emergence).

To analyze one further specific possible cause for why macroscales of biological networks evolved, we calculated how resilience differed for nodes inside of or outside of macro-nodes. We found that resilience of nodes left outside the macro was far lower, on average, than the resilience of nodes grouped into macro-nodes. This indicates that there are direct measurable benefits of having macroscales, such as increased resilience, and that systems with informative macroscales can still have a high effectiveness but also maintain the benefits of having low effectiveness at a microscale. This is in line with the existing research showing that resilience increases with evolution [16].

These findings present evidence that biological systems are sensitive to the tradeoff between effectiveness and robustness by examining whether evolution brings about multiscale structure in biological networks. Systems with a single level of function face an irresolvable paradox: uncertainty in the connections and interactions between nodes leads to resilience to attack and robustness to node failures, but this decreases the effectiveness of that network. However, multiscale systems, defined as those with an informative higher scale, can solve this 'certainty paradox' by having high uncertainty in their connectivity at the microscale while having high certainty in their connectivity at the macroscale. The tradeoffs between being effective at a microscale (typically in prokaryotes, e.g. bacteria) and being noisy at microscale while transitioning the information to higher scales (eukaryota) might have played a key role in evolutionary dynamics. Indeed, the drive from a prokaryotic ancestor to a eukaryotic one might have occurred based on this trade-off, however explaining such a phenomenon is outside the scope of the current work.

It is worth noting that our results on the 'certainty paradox' may be similar to results showing a relationship between evolution and the criticality of biological systems, wherein criticality reflects how close to the 'edge of chaos' a system is. Criticality has been shown in a diverse set of biological systems, such as uniquely distinguishing biological gene regulatory networks [25], or in protein evolution where criticality has been linked to robustness [26], and criticality is even an outcome of evolved artificial neural agents [27]. Although mathematically possessing an informative macroscale is not the same as criticality for a biological system, they both may be important biological properties that drive evolutionary adaptations and are worthy of further investigation.

While we have illuminated many of the advantages of biological macroscales and posited a functional reason for their existence as the solution to the 'certainty paradox,' what are the biological mechanisms behind the evolution of multiscale structure? We offer here a few hypotheses about biological mechanisms that are concordant with the hypothesis of multiscale advantages in terms of having both effectiveness and robustness.

Notably, evolution can proceed both via neutral processes and selection-based contexts. A well-known neutral process that affects interactions at cellular scale, such as those between proteins, is presuppression (also termed constructive neutralism) [20]. This refers to the complexity arising in the dependencies between interacting molecules in the absence of positive selection [19]. Simply put, the likelihood of maintaining independence between partners is less than that of moving away from the original state (by accumulating changes), and therefore random changes can increase the number of interactions between proteins in a system by chance alone and result in 'noisiness' in the interactions. This may offer a biological mechanism behind the result in low effectiveness in an interactome. Because eukaryota have both a larger number of proteins and a higher substitution rate than bacteria [16], eukaryotic interactomes might be expected to feature a higher number of neutral processes, all of which would combine to make interaction networks noisier and less effective. One hypothesis is that neutral evolution specifically drives the noise at the microscale but not the macroscale. At the macroscale, interactomes would be trimmed and evolved under evolutionary constraints and selective pressures [28], which would eventually reinforce beneficial relationships, thinning out those that can cause negative effects on survival or growth [20]. These processes may lead to formation of subgroups of proteins in the network with more and stronger interactions within the group compared with fewer or weaker interactions between those in different subgroups [20, 29]—thereby leading to the emergence of modular, macroscale structures in these networks, which we hypothesize to be correlated with organismal function [11].

Another possible explanation as to the biological mechanism behind our observed results of a decrease in effectiveness is that prokaryotes are more metabolically diverse than eukaryotes, possessing more metabolic processing pathways [30]. Together with changed usage patterns (such as carbon catabolite repression in bacteria), this specificity of metabolite processing reduces energy demand and allows for more effective usage of resources [31]. These processes would make biochemical inputs and outputs more streamlined and efficient in prokaryotes, which in turn should increase the effectiveness of their protein interactomes, given energy and genomic size constraints [32]. In contrast, eukaryota, as a group, are less constrained by energy than prokaryotes [33] but must contend with a constrained number of metabolites, channelizing them to perform cellular functions in morphologically more complex environments [30, 33]. Eukaryotic cells are about three orders of magnitude larger than prokaryotes [33], requiring more and different sets of controls and organizational processes. Prokaryotes depend on free diffusion for intracellular transport, whereas eukaryota have elaborate mechanisms for targeted transfers [34]. This reliance on cellular transport mechanisms can lead to higher modular (and thus more degenerate or indeterministic) structure in protein interactomes and other intracellular entities, which, as we show here, can be associated with less noise at higher scales of interaction. These higher scale inter module transfer mechanisms ensure the proper and less noisy flow of important molecules among these modules (such as protein or metabolite transport among organelles) [11]. Each of these larger scale processes, such as transport among organelles, relies on only a handful of inputs and outputs from outside its module, as compared with much more diverse interactions within the modules themselves [11], which arise due to both functional and neutral processes. In terms of networks, this hierarchical organizational structure is apt to lead to a higher network effectiveness score at the module/process scale compared with the microscale.

Such mechanistic biological explanations for why we might observe these differences in effectiveness are in line with the theoretical reasoning that biological systems need to resolve the paradox they face at individual scales and therefore construct multiscale structure. We seek to tie the 'certainty paradox' directly to the notion of scale in biological systems and provide a means for researchers to reduce the 'black box' nature of these systems by searching across scales for models with low uncertainty. Understanding the mechanics of information transfer and noise in biological systems, and how they affect functionality, remains a major challenge in biology today. One can imagine that the drive from unicellular to multicellular life was based on some form of similar trade-offs, as those between prokaryotes and eukaryotes, that allowed multicellular life to operate via effective macro-states while reserving a pool of noise and degeneracy. Thus, understanding the information structure of these interactomes lends us an eye into the inner workings of long-term evolutionary processes and trade-offs that might have resulted in the two biggest phenotypic splits in evolutionary history—that of prokaryotic and eukaryotic cells, and of unicellular and multicellular life. We hope this developed framework is applied to other interactomes and other biological networks, such as gene regulatory networks, or even functional brain networks, to examine both how uncertainty plays a role in robustness, how informative higher scales change across evolution and what fundamental tradeoffs biological systems face.

METHODS

Protein interactomes

Protein interactomes are complex models of intracellular activity, often based on high-throughput experiments [35, 36]. Here protein interactomes formed from a curated set of high-quality interactions between proteins (PPIs) are taken from the STRING database [14, 15] the curation of which is outlined in Zitnik *et al.* [16]. In this curation, the STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins, found at http://string-db.org) is used to derive a protein interactome for each species. Each PPIs in the protein interactome is an undirected edge where the edges are based on experimentally documented physical interactions in the species itself or on human expert-curated interactions (e.g. no interactions are based on text-mining or associations). The dataset is curated to only include interactions derived from direct biophysical PPIs, metabolic pathway interactions, regulatory protein–DNA interactions and kinase–substrate interactions. The details of the curation of these interactomes can be found in Zitnik *et al.* [16].

The evolutionary history of the set of PPIs was obtained by Zitnik *et al*. [16] and is derived from a high-resolution phylogenetic tree [23]. The tree is composed of archaea, bacteria and eukaryota and captures a diversity of species in each lineage. The phylogenetic tree is used to characterize the evolution of each species based on the total branch length (which takes the form of nucleotide substitutions per site) from the root of the tree to the leaf of the species. The phylogenetic taxonomy, the names of species and lineages of each species were taken from the NCBI Taxonomy database [37]. Details of how this is associated with each species can be found at (http://snap.stanford.edu/tree-of-life), and we refer to Zitnik *et al*. [16] for further specifics on how each species was assigned an average nucleotide substitution rate. Ultimately these protein interactomes are incomplete models that may change as time goes on. Because we do not wish to bias our results, our statistical analyses were performed only over the interactomes of the species based on >1000 citations in the literature.

Spectral analysis for identifying causal emergence

Spectral methods have proved to be successful in identifying good graph partitions in a wide variety of applications [38]. Given an undirected network, we take the degree normalized adjacency matrix A and compute the eigendecomposition $A = E\Lambda E^{\mathrm{T}}$, where the ith column of E is the normalized eigenvector corresponding to the ith eigenvalue, and Λ is the matrix with the ith eigenvalue on the ith diagonal and zeros elsewhere. The eigenvector matrix E contains rich information about the structure of the network, including information about the optimal scale of a network. The rows of E correspond to nodes in the network, so we construct a vector representation of each node's contribution to the network topology by weighting the columns of E by their corresponding eigenvalues, removing columns that correspond to null eigenvalues, and associating the resulting row vectors with the nodes of the network. We construct a distance metric that reflects similarity in causal structure between pairs of nodes by taking the cosine similarity between the vectors corresponding to nodes. If a pair of nodes are not in each other's Markov blankets, coarse graining them together cannot increase the *EI*, so we define the distance between them to be ∞ (or simply very large, in this case, 1000). We use this metric to cluster the nodes of the network using the OPTICS algorithm [39] which we can interpret as a coarse graining to construct a macroscale network, where micro-nodes are placed in the same macro-node if they are placed in the same cluster. Note that this method for detecting causal emergence in networks is explored in detail in other sources [13].

Robustness of causal emergence differences across species

To ensure that the differences observed in the causal emergence values of the PPI networks were not merely a statistical artifact, we conducted a series of robustness tests of our analysis. These tests were necessary for two key reasons. First, the nature of interaction data in biology is inherently difficult to obtain. Although many of the tools we use to collect, clean and interpret biological systems are sophisticated, they are nonetheless subject to potential biases. However, if there were systematic biases in the network construction process for the protein interactomes used in this study (for example, if the interaction networks of eukaryotic species systematically over estimated certain interactions), randomization procedures should clarify the extent to which the results we observed are truly a property of the species themselves.

Second, these robustness tests offer insights into whether there is anything intrinsic to the network structures of the eukaryotic or prokaryotic species that could be contributing to their causal emergence values. For example, the protein-interaction networks of the eukaryote, *Rattus norvegicus* (the common sewer rat), have a certain amount of causal emergence. Would an arbitrary, simulated network with the same number of nodes and edges, connected randomly, also have a similar amount of causal emergence? By performing a series of robustness tests on the protein interaction networks in our study, we can get closer to the question of whether or not there is anything intrinsic to the protein interaction network of *R. norvegicus*, or any other species, that makes it particularly prone to displaying higher scale informative structures?

To address the two concerns above, we performed two separate but similar robustness tests. The first uses a network null model known as the configuration model to randomize the connectivity of the protein interactomes while also preserving the number of nodes, edges and distribution of node degree [40]. The second robustness test involves random edge rewiring [41]. For each network in our study, we iteratively increased the fraction of random edges to rewire in the network; an edge, e_{ij}, that connects nodes v_i and v_j, becomes reconnected to a new node, v_k, forming a new edge, e_{ik}, instead of the original e_{ij}. We do this with an iteratively increasing fraction of edges, starting with 1% of edges and increasing until 100% of the network's edges are rewired.

If the causal emergence values of the networks in this study decrease following the robustness tests above—and in particular, if they decrease differently for eukaryota and prokaryotes—then the differences we observe are unlikely to have arisen simply from chance, noisy/biased data, or otherwise coincidental, *ad hoc* network properties. Instead, our testing of the robustness of our analysis lends credence to the main finding of this paper, which is that species that emerged later in evolutionary time are associated with more informative macroscale protein interaction networks.

In Figure 4A, we show how the causal emergence of archaea, bacteria and eukaryota interactomes all decreases as a higher and higher amount of network

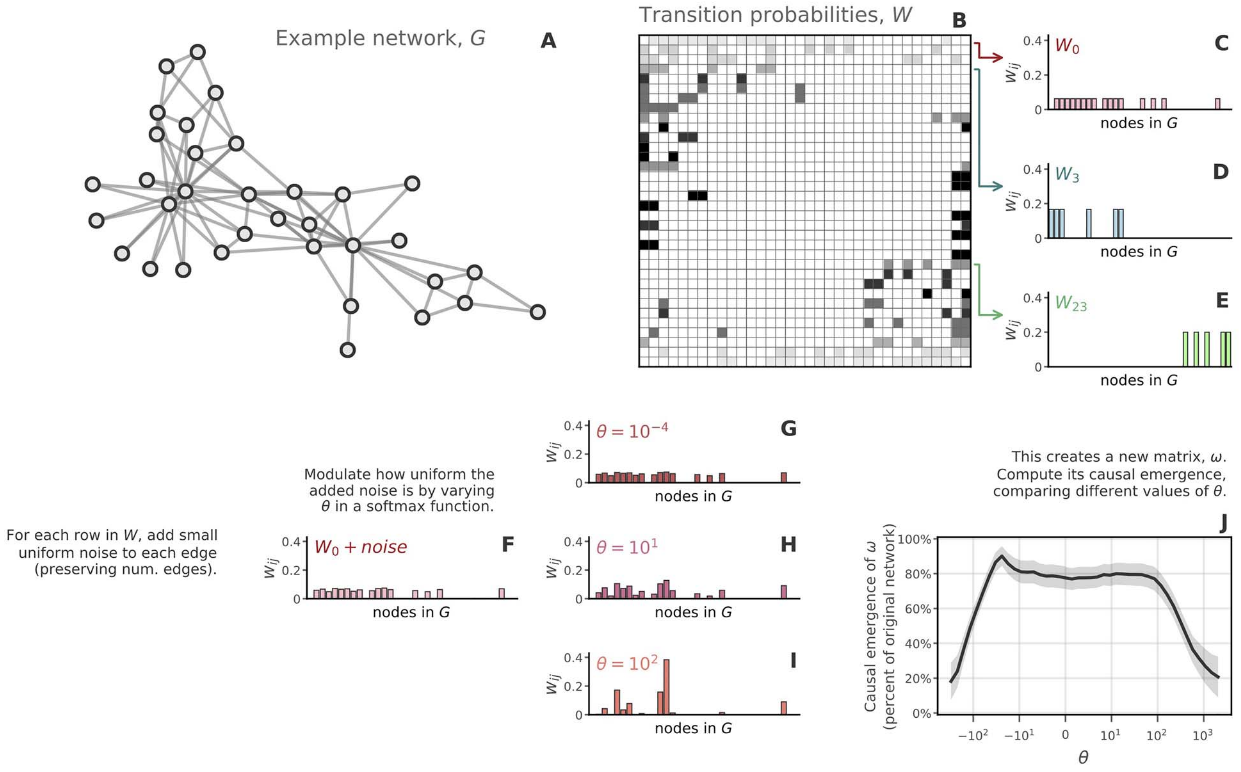
Example network, G
A
Transition probabilities, W
B
W_0
C
W_3
D
W_{23}
E
0.4
0.2
0
w_{ij}
nodes in G
$\theta = 10^{-4}$
G
Modulate how uniform the added noise is by varying θ in a softmax function.
This creates a new matrix, ω. Compute its causal emergence, comparing different values of θ.
For each row in W, add small uniform noise to each edge (preserving num. edges).
$W_0 + noise$
F
$\theta = 10^{1}$
H
J
$\theta = 10^{2}$
I
Causal emergence of ω (percent of original network)
100%
80%
60%
40%
20%
0%
-10^2
-10^1
0
10^1
10^2
10^3
θ

Figure 5. Schematic of edge reweighting procedure. Consider a network, G (A), and its transition probability matrix, W (A). Each element w_{ij} of W corresponds to the probability that a random walker on node v_i will transition to node v_j at the following timestep; for each node v_i, w_{ij} is $1/k_i$, where k_i is the out-degree of node v_i. (C, D, E) Examples of the transition probabilities (i.e. the rows of W) for three nodes in G. The edge reweighting procedure introduced here involves two steps: (F) First, we select a row of W and add small uniform noise to it. Next, we apply a softmax function to the noisy vector, which is defined as $\sigma\left(W_i\right) = \frac{e^{\theta W_i}}{\sum_{j=1}^{n} e^{\theta w_{ij}}}$. In (G, H, I) we show how the output of a softmax function relies on the θ parameter—as θ increases, the slightest differences become accentuated in the resulting value, which we show using the noisy W_0 vector and three different values of θ. (J) We iteratively span values of θ, and we are able to approximate the causal emergence of networks with (tunable) non-uniform edge weights; here we plot the resulting causal emergence values as a percentage of the original causal emergence. Error bands are 95% confidence intervals.

edges are rewired, indicating that random rewiring has a similar effect on all datasets. This analysis suggests that if there were significant noise in the network data itself (i.e. connections between proteins where there otherwise should not be or a lack of connections where there should be), we should not expect to see the magnitude of causal emergence values that we indeed do see. This adds evidence that the inherent noise in the data collection process is not sufficient to produce the results we see.

In Figure 4B, we show that random null models of the networks used in this study are characteristically unlikely to have values for causal emergence values that are at all similar to the original interactomes. On the contrary, the maximum average causal emergence value for any of the networks used here reaches only 3% of the original network's values. This suggests that random null models of networks are less likely to contain higher scale structure but also that the observed differences in the causal emergence values for prokaryotic and eukaryotic species are unlikely to be driven merely due to basic properties like their edge density or degree distribution.

Lastly, *EI* is sensitive not only to the network structure but also to the distribution of edge weights within the network. However, we do not have edge weights data for the particular networks included in this work; as such, we assign uniform weights of $1/k_i$ to the k_i edges of node v_i, which means that edge weights correspond to the probability that a random walker will traverse from node v_i to node v_j in the next time step. This assumption likely distorts the true- (likely non-uniform) weighted interaction patterns between proteins. This could affect the main results of this work, and as such, we devised an edge weight randomization procedure that allows us to systematically vary the shape of the distribution of edge weights in the networks we study. That is, whereas we usually assume that nodes have 'flat' edge weights (uniform, summing to 1.0) to neighboring nodes, under the edge weight randomization, we can create versions of a given network where nodes' edge weights distributions are more or less heavy-tailed. Take, for example, a node connected to four other nodes; originally, its vector of edge weights would be [0.25, 0.25, 0.25, 0.25]. The procedure introduced here allows us to continuously vary the shape of the edge weight distribution to something more heavy-tailed, such as [0.025, 0.05, 0.025, 0.9]. A schematic of the randomization procedure is shown in Figure 5, the results of which are shown in Figure 6

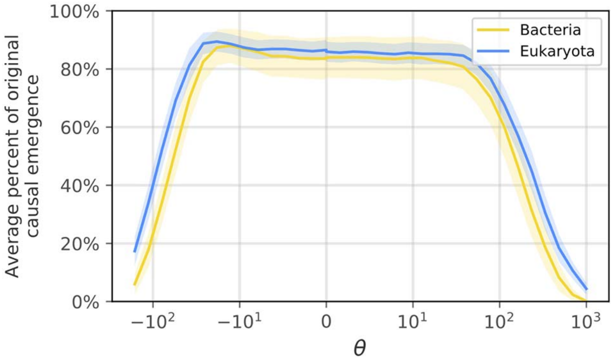

Figure 6. Comparison of causal emergence of bacteria and eukaryota. Using the procedure in Figure 5, we select a random sample of bacteria species and eukaryota species and perform the edge reweighting on their interactomes. Notably, the two domains do not substantially differ under this re-weighting scheme, at least partially validating the results from the main text. Error bands are 95% confidence intervals.

and a full description of the randomization procedure below.

To reweight the edges in a given network, we first multiply each row of W^{out} by vectors of uniformly distributed noise. This multiplication ensures that the vectors of noise will preserve the presence and absence of outgoing edges for each node. After this, we renormalize each row so that they sum to 1.0. Next, we want to introduce tunable non-uniformity to each node's edge weights. For example, if node v_i connects to nodes v_j and v_k each with a weight $w_{ij} = w_{ik} = 0.5$, we want to track how the network's causal emergence changes if these edge weights were, for example, [0.45, 0.55] or [0.7, 0.3] or [0.9, 0.1], etc. For this, we use a *softmax* function—a common tool in machine learning and statistics. Softmax functions have many purposes and can be used to exaggerate or depress probabilities in vectors; the function is defined as:

$$\sigma\left(W_i\right) = \frac{e^{\theta W_i}}{\sum_{j=1}^{N} e^{\theta w_{ij}}} \tag{5}$$

where θ modulates the extent to which higher probability values will become even higher. For example, the vector [0.1, 0.2, 0.3, 0.4], when passed through a softmax function with $\theta = 10$, becomes [0.03, 0.09, 0.24, 0.64]; when $\theta = 0$, it becomes [0.25, 0.25, 0.25, 0.25]. This θ parameter, when varied, can generate a range of uniform and non-uniform vectors—precisely what this edge reweighting procedure requires. We repeatedly create noisy W^{out} matrices and apply the softmax function under a variety of θ values. We can then compute the causal emergence of this new matrix, ω, and calculate its percentage of the original network's causal emergence value. We plot this for an example network under a range of θ values in Figure 5J.

To compare the effect of uniform versus non-uniform edge weights in the protein interaction networks studied in this work, we randomly sample 50 bacteria and 50 eukaryota to perform this procedure on. We compare the effect of this reweighting procedure in Figure 6, finding little differences between bacteria and eukaryota. This result and procedure are important controls for characterizing the higher informative scales of different protein networks, as there does not seem to be especially consistent descriptions of the expected shape of individual proteins' edge weight distributions across species. Note that here we make the minimal assumption that there should not be domain-specific differences between how this reweighting scheme impacts the causal emergence of bacteria versus eukaryota. The introduction of this technique opens a wide number of novel research questions for future work. Importantly, we do not want to introduce additional assumptions about how this non-uniform weight is distributed (e.g. we do not want to artificially impose correlations between edge weight and degree of incident nodes, as we did not find evidence for this in the literature).

Although it is impossible to exhaust all possible sources of bias or confounding variables in biological networks, the two statistical controls performed here get us closer to validating the hypotheses underlying this work: that evolution brings about higher informative scales in protein networks.

Author contributions

E.H., B.K. and M.L. conceived the project. E.H., B.K., A.S. and M.L. wrote the article. E.H., B.K. and R.G. performed the analyses.

Funding

Allen Discovery Center Program through The Paul G. Allen Frontiers Group (12171 to M.L.);Templeton World Charity Foundation, Inc. (TWCFG0273); Army Research Office (proposal 77111-PH-II); National Science Foundation (DGE-1632976 to A.S.). John Templeton Foundation (61780 to B.K. and 62212 to M.L.).

Software and data availability

The dataset and Python code to reproduce these analyses is available at https://github.com/jkbren/einet [42].

Conflict of interest statement

The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation and Templeton World Charity Foundation, Inc..

References

1. Edelman GM, Gally JA. Degeneracy and complexity in biological systems. *Proc Natl Acad Sci U S A* 2001;**98**:13763–8. 10.1073/pnas.231499798.
2. Tsimring LS. Noise in biology. *Rep Prog Phys* 2014;**77**:026601. 10.1088/0034-4885/77/2/026601.

3. Einstein A. On the theory of the Brownian movement. *Ann Phys* 1906;**19**:371–81. https://einsteinpapers.press.princeton.edu/vol2-trans/194.
4. Colquhoun D, Hawkes AG. On the stochastic properties of single ion channels. *Proc R Soc B* 1981;**211**:205–35. 10.1098/rspb.1981.0003.
5. Başar E. *Chaos in Brain Function: Containing Original Chapters by E. Basar and TH Bullock and Topical Articles Reprinted from the Springer Series in Brain Dynamics*. Berlin, Germany: Springer Science & Business Media, 2012.
6. Brennan MD, Cheong R, Levchenko A. How information theory handles cell signaling and uncertainty. *Science* 2012;**338**:334–5. 10.1126/science.1227946.
7. Tononi G, Sporns O, Edelman GM. Measures of degeneracy and redundancy in biological networks. *Proc Natl Acad Sci U S A* 1999;**96**:3257–62. 10.1073/pnas.96.6.3257.
8. Dolinski K, Troyanskaya OG. Implications of big data for cell biology. *Mol Biol Cell* 2015;**26**:2575–8issn: 19394586. 10.1091/mbc.E13-12-0756.
9. Marx V. The big challenges of big data. *Nature* 2013;**498**: 255–60issn: 00280836. 10.1038/498255a.
10. Whitacre JM. Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theor Biol Med Model* 2010;**7**:1–17. 10.1186/1742-4682-7-6.
11. Alon U. Biological networks: the tinkerer as an engineer. *Science* 2003;**301**:1866–7. 10.1126/science.1089072.
12. Bray D. Molecular networks: the top-down view. *Science* 2003;**301**:1864–5. 10.1126/science.1089118.
13. Klein B, Hoel E. The emergence of informative higher scales in complex networks. *Complexity* 2020;**2020**:1–12. 10.1155/2020/8932526.
14. Szklarczyk D, Franceschini A, Kuhn M et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**:561–8. 10.1093/nar/gkq973.
15. Szklarczyk D, Morris JH, Cook H et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–8. 10.1093/nar/gkw937.
16. Zitnik M, Sosič R, Feldman MW et al. Evolution of resilience in protein interactomes across the tree of life. *Proc Natl Acad Sci U S A* 2019;**116**:4426–33. 10.1073/pnas.1818013116.
17. Hoel E, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A* 2013;**110**:19790–5. 10.1073/pnas.1314922110.
18. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13. 10.1038/nrg1272.
19. Lukeš J, Archibald JM, Keeling PJ et al. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 2011;**63**:528–37. 10.1002/iub.489.
20. Brunet TDP, Ford Doolittle W. The generality of constructive neutral evolution. *Biol Philos* 2018;**33**:2. 10.1007/s10539-018-9614-6.
21. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**:379–423. 10.1145/584091.584093.
22. Hoel E. When the map is better than the territory. *Entropy* 2017;**19**:188. 10.3390/e19050188.
23. Hug LA, Baker BJ, Anantharaman K et al. A new view of the tree of life. *Nat Microbiol* 2016;**1**:1–6. 10.1038/nmicrobiol.2016.48.
24. Griebenow R, Klein B, Hoel E. Finding the right scale of a network: efficient identification of causal emergence through spectral clustering. *arXiv* 2019. https://arxiv.org/abs/1908.07565.
25. Daniels BC, Kim H, Moore D et al. Criticality distinguishes the ensemble of biological regulatory networks. *Phys Rev Lett* 2018;**121**:138102. 10.1103/PhysRevLett.121.138102.
26. Tang Q-Y, Hatakeyama TS, Kaneko K. Functional sensitivity and mutational robustness of proteins. *Phys Rev Res* 2020;**2**:033452. 10.1103/PhysRevResearch.2.033452.
27. Khajehabdollahi S, Witkowski O. Evolution towards criticality in Ising neural agents. *Artif Life* 2020;**26**:112–29. 10.1162/artl_a_00309.
28. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 1999;**96**:3801–6. 10.1073/pnas.96.7.3801.
29. Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 2006;**440**:41–5. 10.1038/nature04531.
30. Carlile M. Prokaryotes and eukaryotes: strategies and successes. *Trends Biochem Sci* 1982;**7**:128–30. 10.1016/0968-0004(82)90199-2.
31. Görke B, Stülke J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat Rev Microbiol* 2008;**6**:613. 10.1038/nrmicro1932.
32. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014;**8**:1553. 10.1038/ismej.2014.60.
33. Lane N. Energetics and genetics across the prokaryote-eukaryote divide. *Biol Direct* 2011;**6**:35. 10.1186/1745-6150-6-35.
34. Dacks JB, Peden AA, Field MC. Evolution of specificity in the eukaryotic endomembrane system. *Int J Biochem Cell Biol* 2009;**41**: 330–40. 10.1016/j.biocel.2008.08.041.
35. Rual JF, Venkatesan K, Hao T et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;**437**:1173–8. 10.1038/nature04209.
36. Rolland T, Taşan M, Charloteaux B et al. A proteome-scale map of the human interactome network. *Cell* 2014;**159**:1212–26. 10.1016/j.cell.2014.10.050.
37. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res* 2012;**40**:136–43. 10.1093/nar/gkr1178.
38. Guattery S, Miller GL. On the performance of spectral graph partitioning methods. *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA '95*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1995, 233–242. doi: 10.5555/313651.313702.
39. Ankerst M, Breunig MM, Kriegel H-P *et al*. OPTICS: ordering points to identify the clustering structure. *Proc. ACM SIGMOD'99 Int. Conf. on Management of Data*. Philadelphia, PA, USA: ACM Press, 1999, 49–60. doi: 10.1145/304182. 304187.
40. Garlaschelli D, Loffredo MI. Maximum likelihood: extracting unbiased information from complex networks. *Phys Rev E* 2008;**78**:015101. 10.1103/PhysRevE.78.015101.
41. Karrer B, Levina E, Newman MEJ. Robustness of community structure in networks. *Phys Rev E* 2007;**77**:1–9issn: 1539-3755. 10.1103/PhysRevE.77.046119.
42. Klein B. *jkbren/einet: einet*. Version v1.0. 2021. doi: 10.5281/zenodo.5236550.