1
2

# Evolution leads to emergence: An analysis of protein interactomes across the tree of life

3
4
5
6
7

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

9

10    [1] Allen Discovery Center at Tufts University, Medford, MA, USA

11    [2] Network Science Institute, Northeastern University, Boston, MA, USA

12    [3] Laboratory for the Modeling of Biological & Sociotechnical Systems, Northeastern University,
13    Boston, USA

14    [4] Department of Biology, University of Maryland, College Park, MD, USA

15    [5] Drexel University, Philadelphia, PA, USA

16
17    [*]Author for Correspondence:
18    200 Boston Ave., Suite 4600
19    Medford, MA 02155

20    e-mail: erik.hoel@tufts.edu

21
22    **Running title:**      Evolution leads to higher scales

23
24    **Keywords:**         emergence, information, networks, protein, interactomes, evolution

25
26
27

**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 demonstrate 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 to nodes that do not participate in the macroscale. These effects are more pronounced in interactomes of Eukaryota, as compared to Prokaryota. 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.

**Introduction**

Interactions in biological systems are noisy and degenerate in their functions, making them fundamentally noisier and fundamentally different from those in engineered systems (Edelman & Gally, 2001; Tsimring, 2014). 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 a buffered by Brownian motion (Einstein, 1905), to the stochastic opening and closing of ion channels (Colquhoun & Hawkes, 1981), and even to the chaotic dynamics of neural activity (Başar, 2012).

There are also numerous sources of degeneracy within the cellular, developmental, and genetic operation of organisms (Brennan, Cheong, & Levchenko, 2012). Degeneracy is when an end state or output, like a phenotype, can come from a large number of possible states or inputs networks (Tononi, Sporns, & Edelman, 1999).

Due to this indeterminism and degeneracy, the dynamics and function of biological systems are often uncertain. This hampers control of system-level 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 (Dolinski & Troyanskaya, 2015; Marx, 2013). While there have been many attempts to characterize and understand this uncertainty in biological systems (Tononi, Sporns, & Edelman, 1999), the explanations typically do not extend beyond the advantages of redundancy in these systems (Whitacre, 2010).

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 networks (Alon, 2003; Bray, 2003). 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. In order to quantify this noise and uncertainty we make

76  use of the *effective information* (*EI*), an information-theoretic network quantity based on the
77  entropy of random walker behavior on a network. A lower *EI* indicates greater noise and
78  uncertainty (a formal mathematical definition is given in the Results). Indeed, the *EI* of biological
79  networks has already been shown to be lower in biological networks compared to technological
80  networks (Klein & Hoel, 2020), which opens the question of why this is the case.

81  To see how *EI* changes across evolution, we examined networks of protein-protein
82  interactions (PPIs) from organisms across the tree of life. The dataset consists of *interactomes*
83  from 1840 species (1,539 Bacteria, 111 Archaea, and 190 Eukaryota) derived from the STRING
84  database (Szklarczyk et al., 2010; Szklarczyk et al., 2017). These interactomes have been
85  previously used to study evolution of resilience, where researchers found that species tended to
86  have higher values of *network resilience* with increasing evolution (wherein "evolution" was
87  defined as the number of nucleotide substitutions per site) (Zitnick et al. 2019). In our work, we
88  take a similar approach, highlighting changes in interactome properties as evolution progresses.

89  Additionally, we focus on identifying when interactomes have *informative macroscales*. A
90  *macroscale* refers to some dimension reduction, such as an aggregation, coarse-graining, or
91  grouping, of states or elements of the biological system. In networks this takes the form of
92  replacing subgraphs of the network with individual nodes (*macro-nodes*). A network has an
93  *informative* macroscale when subgraphs of the network can be grouped into macro-nodes such that
94  the resulting dimensionally-reduced network gains *EI* (Klein & Hoel, 2019). When such grouping
95  leads to an increase in *EI*, we describe the resulting macro-node as being part of an informative
96  macroscale. Following previous work, we refer to any gain in *EI* at the macroscale as *causal*
97  *emergence* (Hoel et al., 2013). With these techniques, we can identify which protein-protein
98  interaction (PPI) networks have informative macroscales and which do not. By correlating this
99  property with where (in time) each species lies in the evolutionary tree, we show that informative
100  macroscales tend to emerge *later* in evolution, being associated more with Eukaryota than
101  Prokaryota (such as Bacteria).

102  What is the evolutionary advantage of having informative higher scales? This question is
103  important because higher scales minimize noise or uncertainty in biological networks. Yet such
104  uncertainty or noise represents a fundamental paradox. The more noisy a network is, the more
105  uncertain and the less effective that network is (like being able to effectively transform inputs to
106  outputs, such as being able to effectively upregulate a particular gene in response to a detected
107  chemical in the environment). Therefore, we might expect evolved networks to be highly effective.
108  Yet this is the opposite of what we observe. Instead we observe that effectiveness of lower scales
109  decreases later in evolution, as higher scales that are effective emerge.

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

120  In a biological sense, node failures or attacks in a cellular network may represent certain
121  mutations in proteins or other biochemical entities, which in turn may prevent regular functioning
122  of the system (Barabási & Oltvai, 2004). Biological networks should then, over the course of
123  evolution, develop degeneracy and noise at lower scales to maintain regular functioning, while at

124   the same time developing effectiveness at a higher level. This transformation can be achieved by
125   the action of both neutral and selective processes in evolution. Neutral processes such as pre-
126   suppression, which aided by mutations, increases the number of interactions (Lukeš et al., 2011)
127   and can therefore decrease network effectiveness. On the other hand, selective processes can weed
128   out the noise that interferes with the functioning and efficiency of the system (Brunet & Doolittle,
129   2018). An interplay of these evolutionary processes can lead to a resolution the "certainty paradox"
130   in cellular networks by the develop of informative macroscales.
131         This work therefore presents an explanation for the observed trend in increased resiliency
132   through evolution (Zitnick et al., 2019): informative macroscales make networks more resilient.
133   Finally, we offer insights into biological processes at molecular level that might be responsible for
134   the emergence of informative macroscales in protein-protein interaction networks, specifically
135   looking at the differences between Bacteria, which has a low rate of nucleotide substitutions per
136   site, and Eukaryota, which exhibit a higher rate. Understanding the basic principles governing the
137   differences in efficiency and uncertainty between these major divisions of life can help us
138   comprehend the trade-offs involved in information processes in PPIs across evolution.
139
140   **Results**
141
142   *Effectiveness of protein interactomes across the tree of life.*
143
144   Effective information (*EI*) is a network property reflecting the certainty (or uncertainty) contained
145   in that network's connectivity (Klein & Hoel, 2020). It is a structural property of a network
146   calculated by traversing its topology, and is based on the uncertainty of a random walker's
147   transitions between pairs of nodes, and the distribution of this uncertainty throughout the network.
148   It is calculated by examining the network's connectivity.
149         In a protein interactome, the nodes are individual proteins and the edges of the network are
150   interactions, generally describing the possibility of binding between two proteins. Therefore, the
151   uncertainty we analyze is uncertainty as to which protein(s) a given protein might interact with,
152   e.g., bind with. Each node in the network has out-weights, which are represented by a vector $W_i^{out}$.
153   For instance, protein *A* might share an edge with protein *B* and also protein *C*. Therefore, $W_A^{out}$.
154   is [½, ½]. Since most protein interactome are undirected, its edges are normalized for each node
155   (such that the sum of $W_i^{out}$ for each node is 1.0). Note that this process of normalization implies
156   that the probability of binding is uniform across the different possible interactions. This
157   transformation into a direct network makes the networks amenable to standard tools of network
158   science, such as analyzing random walk dynamics, and it is also necessary to calculate the *EI* of
159   the network. Additionally, the uniform distribution of 1/*n* is the simplest a priori assumption.
160   However, the actual probability of binding is dependent on biological background conditions such
161   as protein prevalence and not included in most open-source models, and therefore our analysis
162   could change if such detailed probabilities were known.
163         The uncertainty associated with each protein can be mathematically captured by examining
164   the entropy of the outputs of a node, $H(W_i^{out})$, wherein a higher entropy indicates more
165   uncertainty more uncertainty as to interactions (Shannon, 1948). The entropy of the distribution of
166   weight across the entire network, $H(\langle W_i^{out} \rangle)$, reflects the spread of uncertainty across the
167   network. A lower $H(\langle W_i^{out} \rangle)$ means that information is distributed only over a small number of
168   nodes. A high $H(\langle W_i^{out} \rangle)$ signifies that information is dispersed throughout the network. The *EI*
169   of a network can then be defined as the entropy of distribution of weights over the network minus
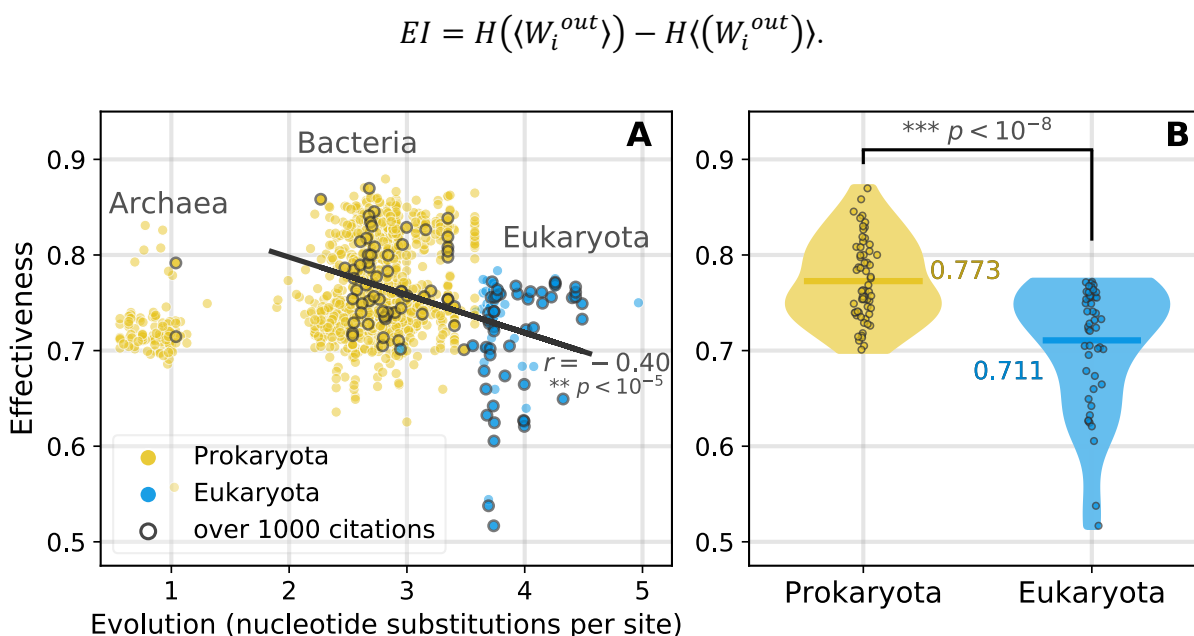170   the average uncertainty inherent in the weight of each node, or:

171
172 $$EI = H\big(\langle W_i{}^{out}\rangle\big) - H\langle \big(W_i{}^{out}\big)\rangle.$$
173



174
175
**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), while 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 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.

184     *EI* can itself be further decomposed into the degeneracy and indeterminism of a network
185 (Klein & Hoel, 2020), where each indicate the lack of specificity in the network's connectivity or
186 interactions. Degeneracy indicates a lack of specificity in targeting nodes (many nodes target the
187 same node), while indeterminism indicates a lack of specificity in targeted nodes (nodes target
188 many nodes). Note that, if networks are considered deterministic in the physical sense, the
189 indeterminism term of *EI* still reflects the uniqueness of targets in the network.
190     A network where all the nodes target a single node will have zero *EI* (since it has maximum
191 degeneracy), as will a network where all nodes target all other nodes (complete indeterminism).
192 *EI* will only be maximal if every node has a unique output. This forces the *EI* of a network to be
193 bounded by $\log_2(n)$. Therefore, in order to compare networks of different sizes, *EI* can be
194 normalized using the number of nodes in the network, *n*, and a new quantity, *effectiveness*, can be
195 defined as:

196 $$\text{effectiveness} = \frac{EI}{\log_2(n)}$$

197
198     To explore the change in effectiveness of biological networks, we examined protein
199 interactomes of 1840 species divided between Archaea, Bacteria, and Eukaryota (see the Methods
200 section for details on the origin and nature of these protein interactomes). We found a clear pattern
201 in the effectiveness of the networks, based on where they are located in the tree of life (Fig. 1), the
202 position of which is based on each protein interactome's small subunit ribosomal RNA gene
203 sequence information (Hug et al., 2016) (see Methods for details). Overall, we found that the mean

204 effectiveness of protein interactomes actually decreases later in the tree of life as nucleotide
205 substitutions occurred. Specifically, Bacteria were found to have a greater effectiveness (0.77)
206 compared to Eukaryota (0.72) on average (student's t-test, $p < 10^{-8}$). Following Zitnik et al.
207 (2019), we restricted further statistical analysis to interactomes with more than 1000 citations, in
208 order use the most well-founded protein interactomes, but the directionality and significance of
209 the result is unchanged when only those above 100 citations are included as compared to when all
210 interactomes are included (student's t-test, $p < 10^{-11}$). Due to the small number of Archaea
211 interactomes based on above 1000 citations we did not include those samples in Fig. 1B.
212
213

214 *Causal emergence across the tree of life.*
215
216 At first the higher effectiveness in Prokaryota interactomes as compared to that of Eukaryota (as
217 shown in Fig. 1) may seem counter-intuitive. One might naively expect the effectiveness of cellular
218 machinery, including or especially interactomes, to increase over evolutionary time, instead of
219 decreasing as we have shown.
220 One hypothesis to explain these results is that, while the protein interactomes get less
221 effective in their micro-scales over evolutionary time, the interactomes are able to nonetheless be
222 effective due to the emergence of informative macroscales as evolution proceeds. To examine this
223 hypothesis, we must first define a procedure for finding macroscales in networks.
224 Network macroscales are defined as subgraphs (i.e., connected sets of nodes and their
225 associated links) that can be grouped into single macro-nodes such that the resulting network has
226 a higher value of *EI* than the original microscale network (Klein & Hoel, 2020). We denote the
227 microscale of a network as $G$ and the macroscale as $G_M$, which is composed of both ungrouped
228 nodes (micro-nodes) and macro-nodes, μ. The macroscale network, $G_M$, is a dimension reduction
229 in that it always has fewer nodes than $G$.
230 To recast a particular subgraph into a macro-node, its connectivity must be modified since
231 the subgraph is being transformed into a single node. In terms of input to the new macro-node, μ,
232 all out-weights that targeted nodes in the subgraph now target the macro-node. In terms of output,
233 each micro-node, $v_i$, inside the subgraph has some $W_i^{out}$. To recast the nodes inside a subgraph
234 into a macro-node, we replace the $W_i^{out}$ of the nodes in the subgraph a single $W_\mu^{out}$, which is a
235 weighted average of the set of each $W_i^{out}$ in the subgraph. The weight is based off the probability
236 $p$ of each node $v_i$ in the stationary distribution of the network, π. This forms macro-nodes (μ|π)
237 that accurately recapitulate the microscale random walk dynamics at the new macroscale (Klein &
238 Hoel, 2020). A macroscale is *informative* if it increases the *EI* of the network compared to the
239 original microscale. In order to find macro-nodes that maximally increase *EI* we make use of a
240 modified spectral algorithm to find locally optimal micro-to-macro mappings, originally described
241 in (Griebenow et al., 2019).
242 Results from this analysis support our initial hypothesis that effectiveness is actually being
243 transitioned to macroscales of biological networks in Eukaryota over evolutionary time, even
244 though the microscales become noisy and less effective over evolutionary time. The total amount
245 of *causal emergence* (the gain of *EI* by grouping subgraphs into macro-nodes) was identified for
246 each protein interactome from each species, normalized by the total size of that protein interactome
247 (Fig 2A). Across the tree of life we observe that Eukaryota have more informative macroscales
248 and show a significant difference in the percentage of microscale nodes that get grouped into
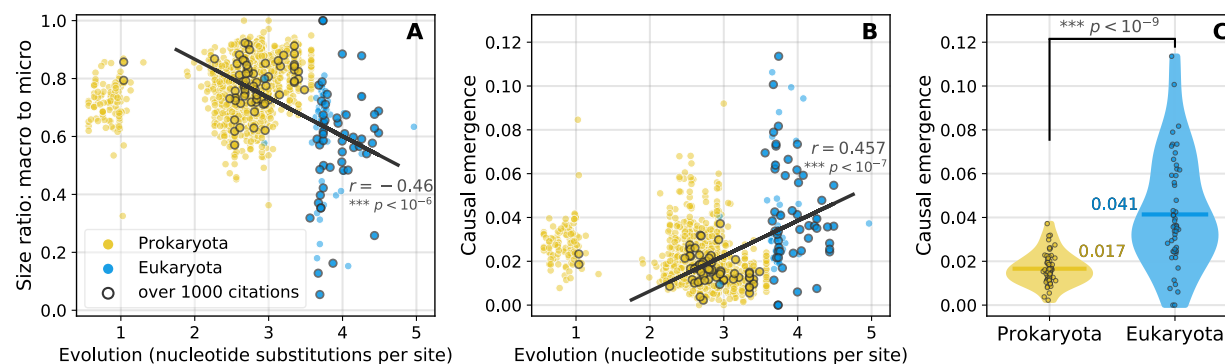249 macro-nodes than (Fig. 2).

**Figure 2**. *Causal emergence in protein interactomes*. **(A)** The protein interactomes of each species undergoes a modified spectral analysis in order 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)** In order 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 ($r = 0.457$, $p < 10^{-7}$). **(C)** The amount of normalized causal emergence is significantly higher for Eukaryota.


*Macroscales of protein interactomes are more resilient than their microscales.*

Why might biological networks evolve over time to have informative macroscales? As previously discussed, one answer might be that having multi-scale structure provides benefits that networks with only a single scale lack. 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.

In order to measure the resilience of the network in response to a node removal we follow (Zitnick et al., 2019) 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:
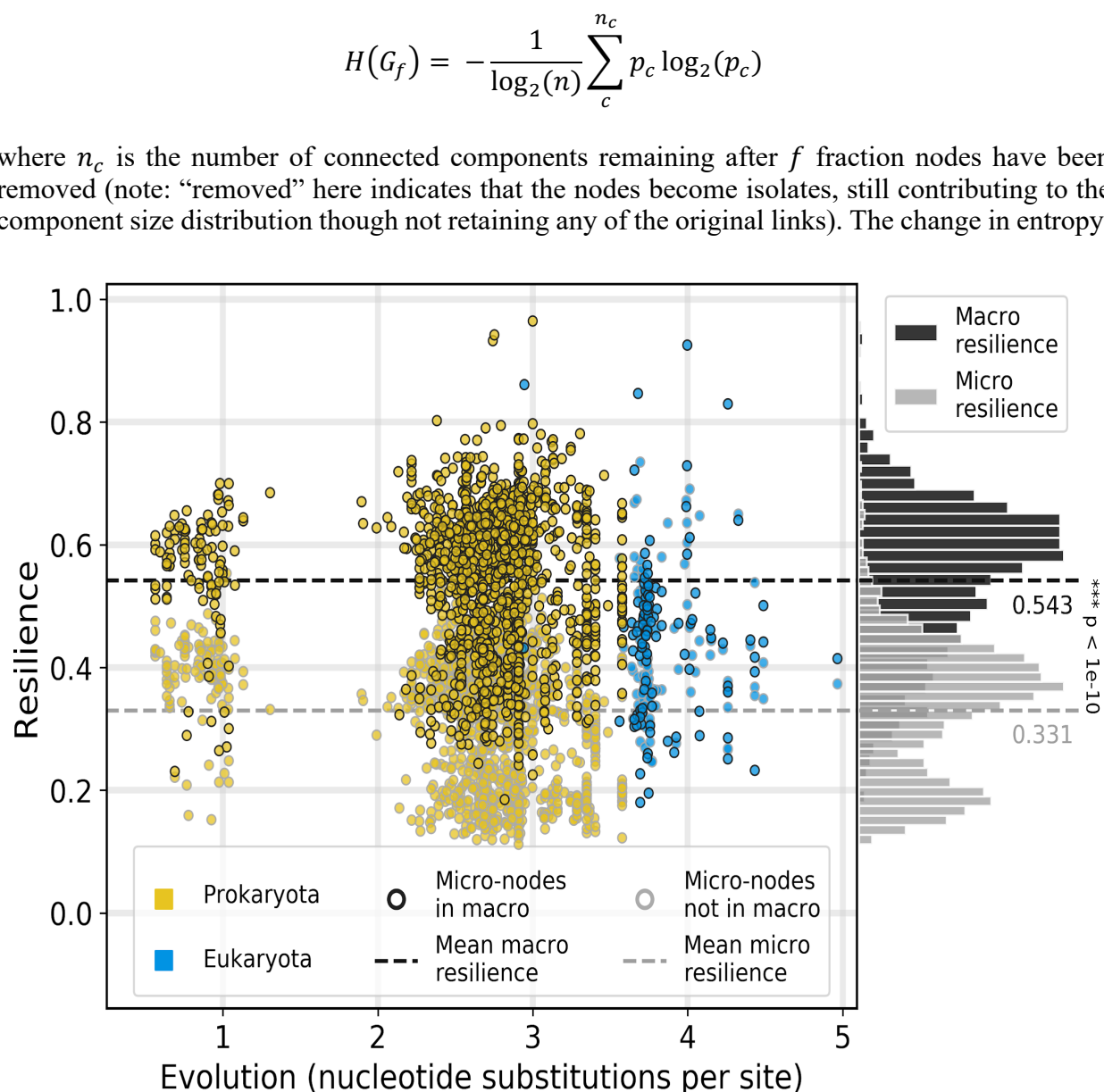
287
$$H(G_f) = -\frac{1}{\log_2(n)} \sum_c^{n_c} p_c \log_2(p_c)$$

288

289  where $n_c$ is the number of connected components remaining after $f$ fraction nodes have been
290  removed (note: "removed" here indicates that the nodes become isolates, still contributing to the
291  component size distribution though not retaining any of the original links). The change in entropy,
292



293
294
295  **Figure 3**. *Resilience of micro- and macro-nodes following causal emergence in interactomes*. The
296  resilience of a species interactome changes across the tree of life, as shown in previous research (Zitnick et
297  al., 2019). Using the mapping generated by computing causal emergence (Fig. 2B), we calculate the
298  resilience of the network, isolating the calculation to nodes that are either part of the macroscale or
299  microscale. Points are color-coded according to the evolutionary domain; points with dark outlines are
300  associated with micro-nodes that have been grouped into a macro-node (macroscale), while the points with
301  light outlines have not been grouped into a macro-nodes (microscale). Nodes at the microscale contribute
302  less to the overall resilience of a given network (0.331) compared to nodes that contribute to macro-nodes
303  (0.543) on average (t-test, $p < 1.0^{-10}$). Note: plotted are the microscale and macroscale resilience values
304  for each interactome in the dataset, and the difference in resilience across scales holds even when only
305  including species with more than 10, 100, or 1000 citations.

306
307

308    $H(G_f)$, as $f$ increases from 0.0 to 1.0 corresponds to the resilience of the network in question.
309    Specifically, this resilience is defined as follows:

310

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

311

312    where $r_f$ is the rate of node removal (i.e., the increment that the fraction $f$ increases from 0.0 to
313    1.0). In this work, we default to a value of $r_f = 100$, which means that the calculation of a
314    network's resilience involves iteratively removing 1%, 2%, ... 100% of the nodes in the network.
315    For each value of $f$, we simulate the node removal process 20 times.
316       Our hypothesis is that biological networks deal with this "certainty paradox" by
317    maintaining uncertainty at their microscale. This gives a pool of noise and degeneracy, leading to
318    resilience. Meanwhile, at the macroscale, the networks can develop a high effectiveness, wherein
319    sets of proteins deterministically and non-degenerately interact. To explore this hypothesis, we
320    compare the network's resilience to removing micro-nodes that are members of subgraphs grouped
321    into macro-nodes to the network's resilience to removing micro-nodes that remain ungrouped
322    (shown in Fig. 3).
323       By isolating the calculation of network resilience to only the micro- or macro-nodes of a
324    network, we see a stark trend emerge wherein nodes inside highly informative macro-nodes are
325    more resilient than nodes outside. That is, nodes in the original interactome that were grouped into
326    a macro-node contribute more to the overall resilience of the interactome. This not only supports
327    our hypothesis that biological networks resolve the "certainty paradox" by building multi-scale
328    structure, but also provides further explanation and contextualization for the recent findings of
329    increasing resilience across evolutionary time (Zitnick et al., 2019).
330
331    **Discussion**
332
333    In this work, we analyzed how the informativeness of protein interactomes changed over
334    evolutionary time. Specifically, we made use of the *effective information* (*EI*) to analyze the
335    amount of uncertainty (or noise) in the connectivity of protein interactomes. We found that the
336    effectiveness (the normalized *EI*) of protein interactomes decreased over evolutionary time,
337    indicating that uncertainty in the connectivity of the interactomes was increasing over evolutionary
338    time. However, we discovered that this was due to eukaryotic protein interactomes possessing
339    higher (informative) scales, such that they had more *EI* when recast as a coarse-grained network—
340    a phenomenon known as causal emergence. This lower effectiveness and higher causal emergence
341    in eukaryotic species was due to the indeterminism and degeneracy in the network structure of
342    their protein-protein interactions.
343       We used a dataset from the STRING database (Szklarczyk et al., 2010; Szklarczyk et al.,
344    2017) that spans more than 1800 species (1,539 Bacteria, 111 Archaea, and 190 Eukaryota), which
345    has been shown to have considerable advantages compared to previous collections of protein
346    interactomes (Zitnik et al., 2019). However, we cannot rule out the possibility that biases might
347    exist in the specific manner of data collection, such as high under-representation of specific types
348    of difficult-to-detect interactions, which could potentially introduce errors in the calculations of
349    effectiveness in eukaryotic interactomes. As such, we conducted a series of statistical robustness
350    tests that accounted for potential biases in both the data collection and network structures of
351    interactomes in our dataset (see Fig. 4 in Methods for further details about these statistical tests).
352    In short, the results we observed in this study cannot be explained by two plausible sources of bias:
353    1) Random rewiring of network edges does not produce similar results and 2) Network null models

354  of each interactome in this study produce only a fraction of the observed causal emergence in our
355  dataset (the maximum causal emergence values for a species' network null model only reached
356  3% of the causal emergence of the original interactome). Notwithstanding these statistical tests, as
357  technology and methods continue to improve these results and hypothesis must be tested rigorous.
358      To analyze why macroscales of biological networks evolved, we calculated how resilience
359  differed for nodes inside of or outside of macro-nodes. We found that resilience of nodes left
360  outside the macro was far lower, on average, than the resilience of nodes grouped into macro-
361  nodes. This indicates that there are benefits of having macroscales, such as increased resilience,
362  and that systems with informative macroscales can still have a high effectiveness but also maintain
363  the benefits of having low effectiveness at a microscale. This is in line with the existing research
364  showing that resilience increases with evolution (Zitnik et al., 2019).
365      These findings present evidence that biological systems are sensitive to the tradeoff
366  between of effectiveness and robustness by examining whether evolution brings about multi-scale
367  structure in biological networks. Systems with a single level of function face an irresolvable
368  paradox: uncertainty in the connections and interactions between nodes leads to resilience to attack
369  and robustness to node failures, but this decreases the effectiveness of that network. However,
370  multi-scale systems, defined as those with an informative higher scale, can solve this "certainty
371  paradox" by having high uncertainty in their connectivity at the microscale while having high
372  certainty in their connectivity at the macroscale. The tradeoffs between being effective at a
373  microscale (typically in prokaryotes, e.g. Bacteria) and being noisy at microscale, while
374  transitioning the information to higher scales (Eukaryota) might have played a key role in
375  evolutionary dynamics. Indeed, the drive from a prokaryotic ancestor to a eukaryotic one might
376  have occurred based on this trade-off, however explaining such a phenomenon is outside the scope
377  of the current work.
378      While we have illuminated many of the advantages of biological macroscales and posited
379  a functional reason for their existence as the solution to the "certainty paradox," what are the
380  biological mechanisms behind the evolution of multi-scale structure? We offer here a few
381  hypotheses about biological mechanisms that are concordant with the hypothesis of multi-scale
382  advantages in terms of having both effectiveness and robustness.
383      Notably, evolution can proceed both via neutral processes and selection-based contexts. A
384  well-known neutral process that affects interactions at cellular scale, such as those between
385  proteins, is pre-suppression (also termed constructive neutralism) (Brunet & Doolittle, 2018). This
386  refers to the complexity arising in the dependencies between interacting molecules in the absence
387  of positive selection (Lukeš et al., 2011). Simply put, the likelihood of maintaining independence
388  between partners is less than that of moving away from the original state (by accumulating
389  changes), and therefore, random changes can increase the number of interactions between proteins
390  in a system by chance alone, and result in "noisiness" in the interactions. This may offer a
391  biological mechanism behind the result in low effectiveness in an interactome. Because Eukaryota
392  have both larger number of proteins and a higher substitution rate than bacteria (Zitnik et al., 2019),
393  eukaryotic interactomes might be expected to feature a higher number of neutral processes, all of
394  which would combine to make interaction networks noisier and less effective. One hypothesis is
395  that neutral evolution specifically drives the noise at the microscale but not the macroscale. At the
396  macroscale interactomes would be trimmed and evolved under evolutionary constraints and
397  selective pressures (Jain, Rivera, & Lake, 1999), which would eventually reinforce beneficial
398  relationships, thinning out those that can cause negative effects on survival or growth (Brunet &
399  Doolittle 2018). These processes may lead to formation of sub-groups of proteins in the network
400  with more and stronger interactions *within* the group compared to fewer or weaker interactions
401  between those in different subgroups (Brunet & Doolittle, 2018; Martin & Koonin, 2006)—

402 thereby leading to the emergence of modular, macroscale structures in these networks, which we
403 hypothesize to be correlated with organismal function (Alon, 2003).
404        Another possible explanation as to the biological mechanism behind our observed results
405 of a decrease in effectiveness is that prokaryotes are more metabolically diverse than eukaryotes,
406 possessing more metabolic processing pathways (Carlile, 1982). Together with changed usage
407 patterns (such as carbon catabolite repression in Bacteria), this specificity of metabolite processing
408 reduces energy demand and allows for more effective usage of resources (Gorke & Stulke, 2008).
409 These processes would make biochemical inputs and outputs more streamlined and efficient in
410 prokaryotes, which in turn, should increase the effectiveness of their protein interactomes, given
411 energy and genomic size constraints (Giovannoni, Thrash, & Temperton, 2014). In contrast,
412 Eukaryota, as a group, are less constrained by energy than prokaryotes (Lane, 2011) but must
413 contend with a constrained number of metabolites, channelizing them to perform cellular functions
414 in morphologically more complex environments (Carlile, 1982; Lane, 2011). Eukaryotic cells are
415 about three orders of magnitude larger than prokaryotes (Lane, 2011), requiring more and different
416 sets of controls and organizational processes. Prokaryotes depend on free diffusion for intracellular
417 transport whereas Eukaryota have elaborate mechanisms for targeted transfers (Dacks, Peden, &
418 Field; 2009). This reliance on cellular transport mechanisms can lead to higher modular (and thus
419 more *degenerate* or *indeterministic*) structure in protein interactomes and other intracellular
420 entities, which, as we show here, can be associated with *less* noise at higher scales of interaction.
421 These higher-scale inter-module transfer mechanisms ensure the proper and less noisy flow of
422 important molecules among these modules (such as protein or metabolite transport among
423 organelles) (Alon, 2003). Each of these larger-scale processes, such as transport among organelles,
424 relies on only a handful of inputs and outputs from outside its module, as compared to much more
425 diverse interactions within the modules themselves (Alon, 2003), which arise due to both
426 functional and neutral processes. In terms of networks, this hierarchical organizational structure is
427 apt to lead to a higher network effectiveness score at the module/process scale compared to the
428 micro-scale.
429        Such mechanistic biological explanations for why we might observe these differences in
430 effectiveness are in line with the theoretical reasoning that biological systems need to resolve the
431 paradox they face at individual scales and therefore construct multi-scale structure. We seek to tie
432 the "certainty paradox" directly to the notion of *scale* in biological systems and provide a means
433 for researchers to reduce the "black box" nature of these systems by searching across scales for
434 models with low uncertainty. Understanding the mechanics of information transfer and noise in
435 biological systems, and how they affect functionality, remains a major challenge in biology today.
436 One can imagine that the drive from unicellular to multicellular life was based on some form of
437 similar trade-offs, as those between prokaryotes and eukaryotes, that allowed multicellular life to
438 operate via effective macro-states while reserving a pool of noise and degeneracy. Thus,
439 understanding the information structure of these interactomes lends us an eye into the inner
440 workings of long-term evolutionary processes and trade-offs that might have resulted in the two
441 biggest phenotypic splits in evolutionary history—that of prokaryotic and eukaryotic cells, and of
442 unicellular and multicellular life. We hope this developed framework is applied to other
443 interactomes and other biological networks, such as GRNs, or even functional brain networks, to
444 examine both how uncertainty plays a role in robustness, how informative higher scales change
445 across evolution, and what fundamental tradeoffs biological systems face.
446
447 **Methods**
448
449 *Protein interactomes.*

450
451  Protein interactomes are complex models of intracellular activity, often based on high-throughput
452  experiments (Rual et al., 2005; Rolland et al., 2014). Here protein interactomes formed from a
453  curated set of high-quality interactions between proteins (protein-to-protein interactions, or PPIs)
454  are taken from the STRING database (Szklarczyk et al., 2010; Szklarczyk et al., 2017) the curation
455  of which is outlined in (Zitnick et al., 2019). In this curation the STRING database (Search Tool
456  for the Retrieval of Interacting Genes/Proteins, found at http://string-db.org) is used to derive a
457  protein interactome for each species. Each PPIs in the protein interactome is an undirected edge
458  where the edges are based on experimentally-documented physical interactions in the species itself
459  or on human expert-curated interactions (e.g., no interactions are based on text-mining or
460  associations). The dataset is curated to only include interactions derived from direct biophysical
461  protein-protein interactions, metabolic pathway interactions, regulatory protein-DNA interactions,
462  and kinase-substrate interactions. The details of the curation of these can interactomes can be found
463  in (Zitnick et al., 2019).
464       The evolutionary history of the set of PPIs was obtained by (Zitnick et al., 2019) and is
465  derived from a high-resolution phylogenetic tree (Hug et al., 2016). The tree is composed of
466  Archaea, Bacteria, and Eukaryota and captures a diversity of species in each lineage. The
467  phylogenetic tree is used to characterize the evolution of each species based on the total branch
468  length (which takes the form of nucleotide substitutions per site) from the root of the tree to the
469  leaf of the species. The phylogenetic taxonomy, the names of species, and lineages of each species
470  were taken from the NCBI Taxonomy database (Federhen, 2011). Details of how this is associated
471  with each species can be found at (http://snap.stanford.edu/tree-of-life), and we refer to (Zitnick et
472  al., 2019) for further specifics on how each species was assigned an average nucleotide substitution
473  rate. Ultimately these protein interactomes are incomplete models that may change as time goes
474  on. Because we do not wish to bias our results, our statistical analyses were performed only over
475  the interactomes of the species based on more than 1000 citations in the literature.
476
477  *Spectral analysis to find macroscales with $EI^{max}$.*
478
479  Spectral methods have proved to be successful in identifying good graph partitions in a wide
480  variety of applications (Guattery & Miller, 1995). Given an undirected network, we take the
481  degree-normalized adjacency matrix $A$ and compute the eigendecomposition $A = E\Lambda E^T$, where
482  the $i^{th}$ column of $E$ is the normalized eigenvector corresponding to the $i^{th}$ eigenvalue, and $\Lambda$ is
483  the matrix with the $i^{th}$ eigenvalue on the $i^{th}$ diagonal and zeros elsewhere. The eigenvector matrix
484  $E$ contains rich information about the structure of the network, including information about the
485  optimal scale of a network. The rows of $E$ correspond to nodes in the network, so we construct a
486  vector representation of each node's contribution to the network topology by weighting the
487  columns of $E$ by their corresponding eigenvalues, removing columns that correspond to null
488  eigenvalues, and associating the resulting row vectors with the nodes of the network. We construct
489  a distance metric that reflects similarity in causal structure between pairs of nodes by taking the
490  cosine similarity between the vectors corresponding to nodes. If a pair of nodes are not in each
491  other's Markov blankets, coarse graining them together cannot increase the effective information,
492  so we define the distance between them to be $\infty$ (or simply very large, in this case, 1000). We use
493  this metric to cluster the nodes of the network using the OPTICS algorithm (Ankerst et al., 1999)
494  which we can interpret as a coarse-graining to construct a macroscale network, where micro-nodes
495  are placed in the same macro-node if they are placed in the same cluster. Note that this method for
496  detecting causal emergence in networks is explore in detail in other sources (Klein & Hoel, 2020).
497

498     *Robustness of causal emergence differences across species interactomes.*

499

500     To ensure that the differences observed in the causal emergence values of the protein-protein
501     interaction networks were not merely a statistical artefact, we conducted a series of robustness
502     tests of our analysis. These tests were necessary for two key reasons. First, the nature of interaction
503     data in biology is inherently difficult to obtain. While many of the tools we use to collect, clean,
504     and interpret biological systems are sophisticated, they are nonetheless subject to potential biases.
505     However, if there were systematic biases in the network construction process for the protein
506     interactomes used in this study (for example, if the interaction networks of eukaryotic species
507     systematically over-estimated certain interactions), randomization procedures should clarify the
508     extent to which the results we observed are truly a property of the species themselves.
509     Second, these robustness tests offer insights into whether there is anything intrinsic to the network
510     structures of the eukaryotic or prokaryotic species that could be contributing to their causal
511     emergence values. For example, the protein interaction networks of the eukaryote, *Rattus*
512     *norvegicus* (the common sewer rat), has a certain amount of causal emergence. Would an arbitrary,
513     simulated network with the same number of nodes and edges, connected randomly, also have a
514     similar amount of causal emergence? By performing a series of robustness tests on the protein
515     interaction networks in our study, we can get closer to the question of whether or not there is
516     anything intrinsic to the protein interaction network of *Rattus norvegicus*, or any other species,
517     that makes it particularly prone to displaying higher-scale informative structures?
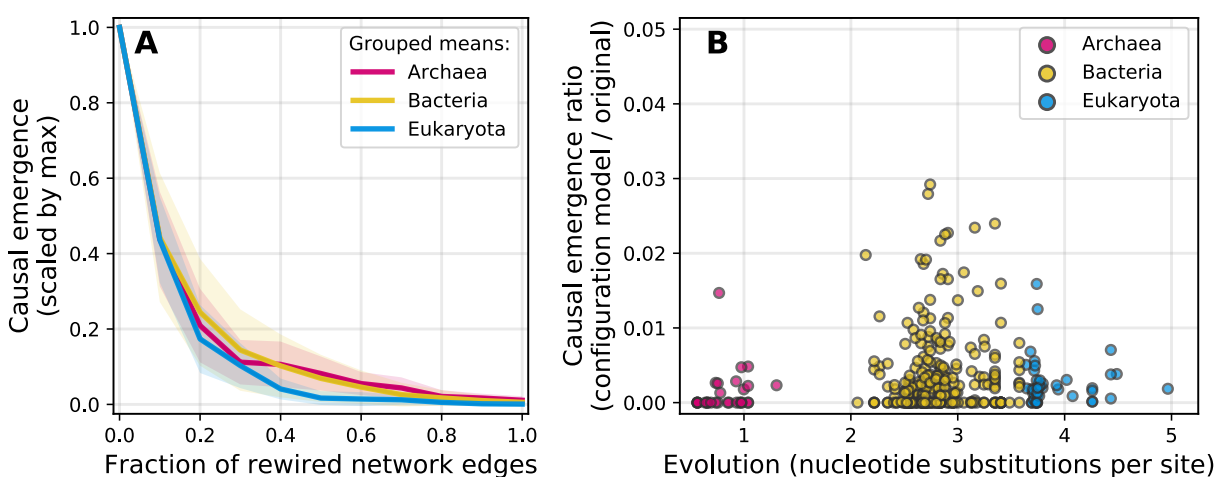
518



519
520

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

530

531        To address the two concerns above, we performed two separate but similar robustness tests.
532     The first uses a network null model known as the *configuration model* in order to randomize the
533     connectivity of the protein interactomes while also preserving the number of nodes, edges, and
534     distribution of node degree (Garlaschelli & Loffredo, 2008). The second robustness test involves

535    random *edge rewiring* (Karrer, Levina, & Newman, 2007). For each network in our study, we
536    iteratively increased the fraction of random edges to rewire in the network; an edge, $e_{ij}$, that
537    connects nodes $v_i$ and $v_j$, becomes re-connected to a new node, $v_k$, forming a new edge, $e_{ik}$, instead
538    of the original $e_{ij}$. We do this with an iteratively-increasing fraction of edges, starting with 1% of
539    edges and increasing until 100% of the network's edges are rewired.
540        If the causal emergence values of the networks in this study decrease following the
541    robustness tests above—and in particular if they decrease *differently* for Eukaryota and
542    prokaryotes—then the differences we observe are unlikely to have arisen simply from chance,
543    noisy/biased data, or otherwise coincidental, *ad hoc* network properties. Instead, our testing of the
544    robustness of our analysis  lend credence to the main finding of this paper, which is that species
545    that emerged later in evolutionary time are associated with more informative macroscale protein
546    interaction networks.
547        In Fig. 4A, we show how the causal emergence of Archaea, Bacteria, and Eukaryota
548    interactomes all decreases as a higher and higher amount of network edges are rewired, indicating
549    that random rewiring has a similar effect on all datasets. This analysis suggests that if there were
550    significant noise in the network data itself (i.e., connections between proteins where there
551    otherwise should not be or a lack of connections where there should be), we should not expect to
552    see the magnitude of causal emergence values that we indeed do see. This adds evidence that the
553    inherent noise in the data collection process is not sufficient to produce the results we see.
554        In Fig. 4B, we show that random null models of the networks used in this study are
555    characteristically unlikely to have values for causal emergence values that are at all similar to the
556    original interactomes. On the contrary, the *maximum* average causal emergence value for any of
557    the networks used here reaches only 3% of the original network's values. This suggests that
558    random null models of networks are less likely to contain higher scale structure but also that the
559    observed differences in the causal emergence values for prokaryotic and eukaryotic species is
560    unlikely to be driven merely due to basic properties like their edge density or degree distribution.
561        While it is impossible to exhaust all possible sources of bias or confounding variables in
562    biological networks, the two statistical controls performed here get us closer to validating the
563    hypotheses underlying this work: that evolution brings about higher informative scales in protein
564    networks.
565
566
567    **Acknowledgements**
568
576
577    **Author contributions**
578
579    E.H., B.K., and M.L. conceived the project. E.H., B.K., A.S. and M.L. wrote the article. E.H.,
580    B.K., and R.G. performed the analyses.
581
582    **Competing interests**

583
584  The authors declare no competing interests.
585
586
587  **References**
588

589  1. Alon, U. (2003). "Biological networks: the tinkerer as an engineer". *Science*, 301(5641),
590     1866-1867.

591  2. Ankerst, M., Breunig, M., Kriegel, H.P., & Sander, J. (1999). "OPTICS: ordering points to
592     identify the clustering structure". *ACM Sigmod Record*, ACM.

593  3. Barabási, A.L., & Oltvai, Z.N. (2004). "Network biology: Understanding the cell's
594     functional organization". *Nature Reviews Genetics*, 5(2), 101.

595  4. Başar, E. (2012). Chaos in Brain Function: Containing Original Chapters by E. Basar and
596     TH Bullock and Topical Articles Reprinted from the Springer Series in Brain Dynamics,
597     Springer Science & Business Media.

598  5. Bizzarri, M., Brash, D., Briscoe, J., Grieneisen, V., Stern, C., & Levin, M. (2019). "A call
599     for a better understanding of causation in cell biology". *Nature Reviews Molecular Cell
600     Biology*, 20(5): 261.

601  6. Bizzarri, M., Palombi, A., & Cucina, A. (2013). "Theoretical aspects of systems biology".
602     *Progress in Biophysics and Molecular Biology,* 112(1-2): 33-43.

603  7. Bray, D. (2003). "Molecular networks: the top-down view". *Science*, 301, 1864–1865.

604  8. Brennan, M.D., Cheong, R., & Levchenko, A. (2012). "How information theory handles
605     cell signaling and uncertainty". *Science*, 338(6105), 334-335.

606  9. Brunet, T.D.P., & Doolittle, W.F. (2018). "The generality of constructive neutral
607     evolution". *Biology & Philosophy*, 33(1-2), 2.

608  10. Carlile, M. (1982). "Prokaryotes and eukaryotes: strategies and successes". *Trends in
609      Biochemical Science*, 7: 128-130.

610  11. Colquhoun, D., & Hawkes, A. (1981). On the stochastic properties of single ion channels.
611      *Proceedings of the Royal Society of London. Series B. Biological Sciences, 211*(1183), 205-
612      235.

613  12. Dacks, J.B., Peden, A.A., & Field, M.C. (2009). "Evolution of specificity in the eukaryotic
614      endomembrane system". *International Journal of Biochemistry & Cell Biology*, 41: 330–
615      340.

616  13. Dolinski, K., & Troyanskaya, O. G. (2015). Implications of Big Data for cell
617      biology. *Molecular biology of the cell*, *26*(14), 2575-2578.

618  14. Edelman, G. M. & Gally, J. A. (2001). "Degeneracy and complexity in biological systems".
619      *Proceedings of the National Academy of Sciences*, 98(24): 13763-13768.

620  15. Federhen, S. (2011). "The NCBI taxonomy database". *Nucleic Acids Research*, 40(D1):
621      D136-D143.

622  16. Garlaschelli, D., Loffredo, M. (2008). "Maximum likelihood: Extracting unbiased
623      information from complex networks". *Physical Review E*, 78.1, 015101.

624  17. Einstein, A. (1905). "Investigations on the Theory of the Brownian Movement." Ann. der
625      Physik.

18. Giovannoni, S.J., Thrash, J.C., & Temperton, B. (2014). "Implications of streamlining theory for microbial ecology". *The ISME Journal*, 8: 1553.

19. Görke, B. & Stülke, J. (2008). "Carbon catabolite repression in bacteria: many ways to make the most out of nutrients". *Nature Reviews Microbiology*, 6: 613.

20. Griebenow, R., Klein, B., & Hoel, E. (2019). "Finding the right scale of a network: Efficient identification of causal emergence through spectral clustering". arXiv preprint arXiv:1908.07565.

21. Guattery, S. & Miller, G. (1995). "On the performance of spectral graph partitioning methods". In: *SODA*.

22. Hoel, E., Albantakis, L., & Tononi, G. (2013). "Quantifying causal emergence shows that macro can beat micro". *Proceedings of the National Academy of Sciences*, 110(49): 19790-19795.

23. Hoel, E. (2017). "When the map is better than the territory". *Entropy*, 19(5):188.

24. Hug, L.A., Baker, B., Anantharaman, K., Brown, C., Probst, A., Castelle, C., Butterfield, C., Hernsdorf, A., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D., Finstad, K., Amundson, R., Thomas, B., & Banfield, J. (2016). "A new view of the tree of life". *Nature Microbiology*, 1(5): 16048.

25. Jain, R., Rivera, M.C., & Lake, J.A. (1999). "Horizontal gene transfer among genomes: the complexity hypothesis". *Proceedings of the National Academy of Sciences*, 96: 3801–3806.

26. Karrer, B., Levina, E., & Newman, M. E. J. (2007). "Robustness of community structure in networks." *Physical Review E*, 77(4), 1–9.

27. Klein, B., & Hoel, E. (2020). The Emergence of Informative Higher Scales in Complex Networks. *Complexity*, *2020*.

28. Lane, N. (2011). "Energetics and genetics across the prokaryote-eukaryote divide". *Biology Direct*, 6(1), 35.

29. Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., & Gray, M.W. (2011). "How a neutral evolutionary ratchet can build cellular complexity". *IUBMB Life*, 63: 528-537

30. Markram, H. (2006). "The blue brain project". *Nature Reviews Neuroscience*, 7(2): 153.

31. Martin, W., & Koonin, E.V. (2006). "Introns and the origin of nucleus–cytosol compartmentalization". *Nature*, 440(7080), 41.

32. Marx, V. (2013). Biology: The big challenges of big data. In: Nature Publishing Group.

33. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S.D., Yang, X., Ghamsari, L., Balcha, D., Begg, B.E., Braun, P., Brehme, M., Broly, M.P., Carvunis, A.R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab F., Gutierrez, B.J., Hardy, M.F., Jin, M., Kang, S., Kiros, R., Lin, G.N., Luck K., MacWilliams, A., Menche, J., Murray, R.R., Palagi, A., Poulin, M.M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J.M., Scholz, A., Shah, A.A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A.O., Trigg, S.A., Twizere, J.C., Vega, K., Walsh, J., Cusick, M.E., Xia, Y., Barabási, A.L., Iakoucheva, L.M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M.A., Hill, D.E.,

668  Hao, T., Roth, F.P., & Vidal, M. (2014). "A proteome-scale map of the human interactome
669  network". *Cell*, 159(5): 1212-1226.

670  34. Rual J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F.,
671  Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M.,
672  Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S.,
673  Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski,
674  R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-
675  Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M. (2005). "Towards a proteome-
676  scale map of the human protein–protein interaction network". *Nature*, 437(7062): 1173.

677  35. Shannon, C.E. (1948). "A mathematical theory of communication". *Bell System Technical
678  Journal*, 27(3): 379-423.

679  36. Szklarczyk D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks,
680  T., Stark, M., Muller, J., Bork, P., Jensen, L.J., von Mering, C. (2011). "The STRING
681  database in 2011: functional interaction networks of proteins, globally integrated and
682  scored". *Nucleic Acids Research*, 39(suppl_1): D561-D568.

683  37. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A.,
684  Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C. (2017). "The STRING
685  database in 2017: quality-controlled protein–protein association networks, made broadly
686  accessible". *Nucleic Acids Research*, gkw937.

687  38. Tononi, G., Sporns, O., & Edelman, G.M. (1999). "Measures of degeneracy and
688  redundancy in biological networks". *Proceedings of the National Academy of Sciences*,
689  96(6), 3257–3262.

690  39. Tsimring, L.S. (2014). "Noise in biology". *Reports on Progress in Physics*, 77(2): 026601.

691  40. Whitacre, J.M. (2010). "Degeneracy: a link between evolvability, robustness and
692  complexity in biological systems". *Theoretical Biology and Medical Modelling*, 7(1): 6.

693  41. Zitnik, M., Sosič, R., Feldman, M.W., & Leskovec, J. (2019). "Evolution of resilience in
694  protein interactomes across the tree of life". *Proceedings of the National Academy of
695  Sciences*, 116(10), 4426–4433.