Universal Scaling in Biochemical Networks

Hyunju Kim[1,2,†], Harrison B. Smith[2,†], Cole Mathis[1,3], Jason Raymond[2] and Sara I. Walker[1,2,4*]

1. Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe AZ USA.
2. School of Earth and Space Exploration, Arizona State University, Tempe AZ USA.
3. Department of Physics, Arizona State University, Tempe AZ USA.
4. Blue Marble Space Institute of Science, Seattle WA USA.

† These authors contributed equally to this work.
* To whom correspondence should be addressed. Email: sara.i.walker@asu.edu.

Abstract: Attempts to identify universal properties of life are limited by the single example of biochemistry on Earth. Using a global database of 28,146 annotated genomes and metagenomes, we report universal scaling laws governing the topology of biochemical networks across levels of organization from individuals, to ecosystems, to the biosphere as a whole. We show the three domains of life are topologically distinguishable, while nonetheless conforming to the same universal scaling laws. Comparing real biochemical networks to networks composed of randomly sampled biochemical reactions reveals the observed scaling is not a product of shared biochemistry alone, but is instead attributable to universal constraints on how global biochemistry is partitioned into individuals. The reemergence of the same regularities across levels of organization hints at general principles governing biochemical network architecture.

There is increasing interest in whether life has universal properties, not tied to its specific chemical instantiation (*1*). Universal biology, if it exists, would have important implications for constraining the origins of life, engineering synthetic life and guiding the search for life beyond Earth (*2*, *3*). Systems biology provides promising tools for uncovering such general principles. One approach has been to study the topology of organismal metabolic networks, revealing common 'scale-free' structure for all three domains of life (*4*). Another approach is identification of scaling laws describing remarkable consistency in scaling of biomass, biodiversity, and metabolic rate (*5–9*). However, so far it has been difficult to unify these patterns across all levels of organization where living processes persist, from individuals to ecosystems to the biosphere. A more convincing case for universal biology would be made by demonstrating all life on Earth shares similar properties, independent of biochemical diversity or level of organization. Here we show biochemical networks conform to universal scaling laws governing their global topology, which tightly constrain not only the network structure of individuals, but also that of ecosystems and the biosphere as a whole.

To study biochemical networks across levels of organization we leverage available annotated genomic and metagenomic data, including genomes of 21,637 bacterial taxa and 845 archaeal taxa from the Pathosystems Resource Integration Center (PATRIC) (*10*), and 77 eukaryotic taxa and 5587 metagenomes from the Joint Genome Institute (JGI) (*11*). From this data, we constructed biochemical networks for each individual organism (genome) and ecosystem (metagenome) using reaction data cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (*12*). Reactions are encoded in our network representation with reactants and products connected to one another

(Supplement methods). Previous topological analyses of biochemical networks have primarily focused on the subset of biochemical reactions associated with metabolism (*12–14*). Since we are interested in properties universal across life, and not just subsets of living processes, we instead construct networks inclusive of every known catalyzed reaction (regardless of pathway) coded by the respective genome or metagenome, provided the reaction is cataloged in KEGG (*15*). We also constructed a network comprising all enzymatically catalyzed reactions cataloged in the KEGG database as a proxy for the biosphere (*15*, *16*).

Adopting a network representation allows systematic quantification of global topological properties using graph theory and statistical mechanics (*13*, *14*, *17*, *18*). We first analyzed degree distributions following methods outlined in (*19*, *20*) to compare candidate fits across levels of organization (Supplement methods). We find >98% of individual level biochemical networks exhibit degree distributions plausibly described by a power-law, based on normalized log-likelihood ratio comparisons to other distributions (Table S2). Power-law fits ($P(k) = c\ k^{-\alpha}$) yield exponents $2 < \alpha < 3$ (Fig. S2), consistent with previously reported values for the subnetworks associated with organismal metabolism (*13*). Surprisingly we find ecosystem level networks can also plausibly be described by a power law when compared to other distributions, with scaling exponents similar to individuals. Since in general it is not true subnetworks of scale-free networks are also scale-free (*21*), this indicates biochemistry has been optimized across levels to yield a common topological structure, potentially reflective of universal constraints governing biochemical network architecture. However, it is well-known degree distribution alone cannot tightly constrain network topology (*14*, *22*, *23*).

To determine whether universal constraints do indeed underlie biochemical network structure, we therefore sought to identify whether scaling laws apply across networks and across levels of organization. To do so, we calculated several frequently implemented topological measures to determine if, and if so how, topology scales with biochemical network size (described in Supplement). Measured values are shown in Fig. 1, plotted as a function of the number of compounds (size) in the largest connected component. We find individuals and ecosystems scale according to the same preferred fit function for each network measure, with similar scaling coefficients (for fits and confidence intervals, see Table S5). Degree assortativity and attribute assortativity were found to be independent of size, with respective means of -0.21 and 0.004 for individuals, and -0.23 and -0.002 for ecosystems. Average betweenness, average shortest path length, number of enzyme classes (a proxy for enzymatic diversity), number of reactions, and number of edges all follow power-law fits ($y = y_0\ x^{\beta}$) with scaling exponents (β) -1.1581, -0.117, 1.294, 1.229, 1.219, respectively for individuals and -1.136, -0.084, 1.838, 1.319, 1.243 respectively, for ecosystems. Average clustering coefficient and average degree follow linear fits ($y = mx + y_0$) with slopes (m) of 3.77×10^{-5} and 6.796×10^{-4} respectively for individuals, and 3.32×10^{-5} and 5.452×10^{-4} respectively for ecosystems. For several topological measures (number of reactions, average shortest path length, and both assortativity and betweenness measures) the biosphere falls within the 95% confidence interval observed for fits of ecosystem level scaling. These results demonstrate biochemical network topology indeed scales with size, with consistent scaling behavior observed for individuals, ecosystems and the entire biosphere.

The measured differences in scaling coefficients for individuals and ecosystems could be statistically insignificant or representative of similar constraints re-emerging at different levels.

We therefore next sought to determine whether or not scaling behavior for individuals is statistically distinguishable from ecosystems. We assumed as a null hypothesis scaling relationships consistent across levels of organization and performed a permutation test (*24*), using the scaling coefficient as the test statistic (SI methods). We find scaling relationships are not distinguishable for individuals and ecosystems when analyzing average node betweenness and average shortest path length. However, scaling relationships are distinguishable for number of reactions, number of edges, number of enzyme classes, mean degree, and mean clustering coefficient, with p-values $< 10^{-5}$ in most cases. Confidence intervals on scaling coefficients for ecosystem topology are narrower than for individuals, indicating ecosystem scaling is more tightly constrained. Although biochemical networks for individuals and ecosystems share similar scaling behavior, they are not drawn from the same distributions; allowing the possibility shared constraints operate at each level separately.

The most deeply rooted division of life is among the three domains, each representing significantly different metabolic strategies and genetic architectures(*25*). If there are universal constraints on biochemical network architecture, as our results indicate, any variation between domains must exist within those constraints. To confirm this, we tested how accurately topological information can predict domain (Supplement methods). We find in most cases global topology normalized to size reliably predicts domain, shown in Table 1, in many cases with > 90% accuracy, whereas topology or size alone are insufficient to make accurate predictions. This confirms the three domains are topologically distinct despite conforming to the same universal scaling relationships. One possible explanation is the three domains utilize different chemical compounds in their biochemistry: however, only a small subset of compounds across our dataset are unique to each domain (Fig. 2, caption), and the most highly connected nodes (compounds participating in many reactions) are shared across all three domains (*13*). This indicates topological differences between the domains are not driven by differences in use of chemical compounds, but rather by how universal constraints dictate the way topology varies with the diversity of compounds (quantified by size of largest connected component) used by individual organisms, with eukaryotes, on average containing the most biochemical diversity and archaea the least diversity.

Given the three domains are topologically distinct within the constraints of universal scaling, a question of interest is whether the observed differences in scaling coefficients for ecosystems as compared to individuals is an emergent property arising from the combination of the three domains. To address this, we generated ecosystem-level networks by merging randomly sampled genome networks from each domain individually and from all three domains together, to determine how scaling behavior could be the same or different for an archeasphere (archaea alone), bacteriasphere (bacteria alone), eukaryasphere (eukarya alone), or artificial ecosystems (all three domains) (Fig. 3). We find the preferred fit functions are the same for real ecosystems and randomly merged organismal networks (random genome networks). Given it is not in general true randomly merging networks produces networks with similar topology, this indicates global constraints on the structure of individuals necessary to universally produce similar scaling behavior in ecosystems. We find however the scaling exponents and coefficients while similar, are nonetheless statistically distinguishable between real ecosystems and random genome networks (see Table S6). Like individual level networks, random genome networks and real ecosystems exhibit exponents distinguishing their scaling relationships for most topological measures and for enzymatic diversity with p-values $< 10^{-5}$. Scaling of betweenness is indistinguishable between the

two datasets. These results indicate random genome networks differ from real ecosystems in many of the same ways individuals do. However, scaling of assortativity does distinguish random genome networks from real ecosystems, whereas it does not distinguish individuals from ecosystems. Taken together, these results suggest scaling behavior for ecosystems arises due to constraints re-emerging at the level of ecosystems.

All individuals share a common subset of compounds (Fig. 2)(*26*), allowing the possibility the observed scaling emerges solely due to biochemistry universally shared across all life on Earth. To test the hypothesis the observed topological scaling emerges due to the presence of universal biochemical compounds (ATP, H_2O etc.), we constructed a different type of reaction network by merging randomly sampled reactions from the KEGG database (random reaction networks), generating networks composed of biochemical reactions but with no notion of individuals. Because most highly connected (participate in many reactions) nodes are common to all three domains, this uniform sampling procedure yields random networks that tend to include the most common compounds used by life. Like real biochemical networks, we find the majority of random reaction networks (98.84%) have heavy-tailed degree distributions (Table S3), plausibly fit to a power-law with $2 < \alpha < 3$: any differences with real biochemical systems therefore cannot be derived from degree distribution alone. Unlike random genome networks, scaling behavior for random reaction networks does not share a preferred fit function with biological data for most network measures. Fits for average clustering coefficient and average degree of random reaction networks favor a power-law function, compared to the linear function favored by biological networks. Fits for number of edges, number of reactions, and attribute assortativity favor a linear function, compared to the power-law, or zeroth-order fit functions favored by biological networks (Table S5). Given these differences in scaling behavior, we conclude shared biochemistry across organisms cannot alone explain the observed scaling laws, the constraints imposed by individuals as a specific partitioning of the biosphere-level network is necessary.

Our analysis reveals the topology of biochemical networks displays common scaling behavior, independent of evolutionary domain or level of organization. Recurrent patterns in different 'units' (individuals, ecosystems, biosphere) at different levels is suggestive of the same basic processes re-emerging across levels of organization. Our results indicate ecosystems are more tightly constrained than individuals, better displaying the regularities of biochemical network architecture. This combined with differences in scaling for random reaction networks reveals an important role for individuals as partitions within the biosphere-level network, necessary to achieve the scaling behavior characterizing life on Earth. One important implication for the emergence of life as a planetary scale transition (*27*) is the necessity of individuality for mediating the transition in the global topological organization of prebiotic chemistry from that of random reaction networks to the scaling reported here. Allometric scaling laws are derived by viewing living systems as localized physical *objects* with energy and power constraints. Here, scaling emerges due to an orthogonal view of living systems as distributed *processes* transforming matter within the space of chemical reactions. It remains an open question what constraints might explain this structure, and whether we should expect all life to exhibit similar scaling behavior.

References

1. N. Goldenfeld, C. Woese, Biology's next revolution. *Nature*. **445**, 369 (2007).

2. D. J. Des Marais *et al.*, The NASA Astrobiology Roadmap. *Astrobiology*. **8**, 715–730 (2008).

3. L. Cronin, S. I. Walker, ORIGIN OF LIFE. Beyond prebiotic chemistry. *Science*. **352**, 1174–1175 (2016).

4. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabási, The large-scale organization of metabolic networks. *Nature*. **407**, 651–654 (2000).

5. G. B. West, J. H. Brown, B. J. Enquist, A general model for the origin of allometric scaling laws in biology. *Science*. **276**, 122–126 (1997).

6. G. B. West, J. H. Brown, The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *J. Exp. Biol.* **208**, 1575–1592 (2005).

7. K. J. Locey, J. T. Lennon, Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5970–5975 (2016).

8. I. A. Hatton *et al.*, The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes. *Science*. **349**, aac6284 (2015).

9. J. P. DeLong, J. G. Okie, M. E. Moses, R. M. Sibly, J. H. Brown, Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12941–12945 (2010).

10. A. R. Wattam *et al.*, Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **45**, D535–D542 (2017).

11. V. M. Markowitz *et al.*, IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* **40**, D123–9 (2012).

12. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

13. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabási, The large-scale organization of metabolic networks. *Nature*. **407**, 651–654 (2000).

14. E. Ravasz, Hierarchical Organization of Modularity in Metabolic Networks. *Science*. **297**, 1551–1555 (2002).

15. J. Raymond, D. Segrè, The effect of oxygen on biochemical networks and the evolution of complex life. *Science*. **311**, 1764–1767 (2006).

16. J. E. Goldford, H. Hartman, T. F. Smith, D. Segrè, Remnants of an Ancient Metabolism without Phosphate. *Cell*. **168**, 1126–1134.e9 (2017).

17. R. Albert, Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).

18. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

19. A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-Law Distributions in Empirical Data. *SIAM Rev.* **51**, 661–703 (2009).

20. J. Alstott, E. Bullmore, D. Plenz, Powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS One*. **9**, e85777 (2014).

21. M. P. H. Stumpf, C. Wiuf, R. M. May, Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4221–4224 (2005).

22. H. Kim, Z. Toroczkai, P. L. Erdős, I. Miklós, L. A. Székely, Degree-based graph construction. *J. Phys. A: Math. Theor.* **42**, 392001 (2009).

23. R. Khanin, E. Wit, How scale-free are biological networks. *J. Comput. Biol.* **13**, 810–818 (2006).

24. S. T. Buckland, B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap. *Biometrics*. **50**, 890 (1994).

25. L. A. Hug *et al.*, A new view of the tree of life. *Nat Microbiol*. **1**, 16048 (2016).

26. N. R. Pace, The universal nature of biochemistry. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 805–808 (2001).

27. E. Smith, H. J. Morowitz, *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere* (Cambridge University Press, 2016).

28. Q. H. Vuong, Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*. **57**, 307–333 (1989).

29. S. Asmussen, *Applied Probability and Queues* (Springer Science & Business Media, 2008).

30. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R* (Springer Science & Business Media, 2013).

31. M. Newman, *Networks: An Introduction* (Oxford University Press, 2010).

32. A.-L. Barabási, *Network Science* (Cambridge University Press, 2016).

33. A. Hagberg, P. Swart, D. S Chult, "Exploring network structure, dynamics, and function using NetworkX" (Los Alamos National Laboratory (LANL), 2008), (available at http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495).

34. M. E. J. Newman, Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).

35. M. E. J. Newman, Mixing patterns in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 026126 (2003).

36. U. Brandes, A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**, 163–177 (2001).

37. U. Brandes, On variants of shortest-path betweenness centrality and their generic computation. *Soc. Networks*. **30**, 136–145 (2008).

38. U. Brandes, C. Pich, Centrality Estimation in Large Networks. *Int. J. Bifurcat. Chaos*. **17**, 2303–2318 (2007).

Acknowledgements: We thank the Emergence@ASU team for feedback through various stages of this work and are grateful for support from the National Aeronautics and Space Administration through grant NNX15AL24G S02, which made this research possible.

Figures and Tables

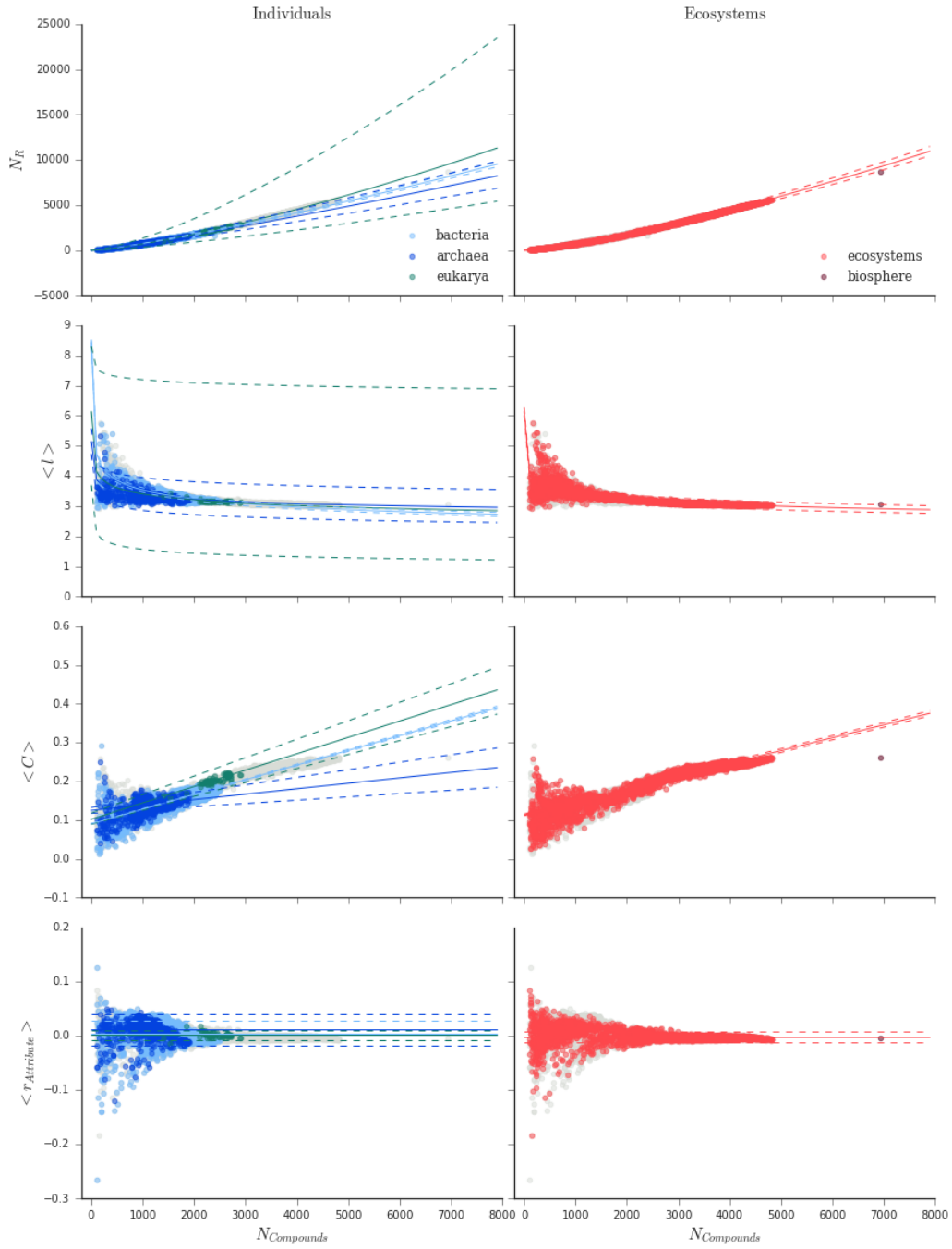

Fig. 1: Scaling of biochemical network topology for individuals (left column) and ecosystems (right column). Shown from top to bottom are number of reactions (N_R), average shortest path ($<l>$), average clustering coefficient ($<C>$), and attribute assortativity ($<r_{Attribute}>$) as a function of the size of the largest connected component ($N_{compounds}$). Individuals include bacteria (light blue), archaea (dark blue), and eukarya (blue-green). Ecosystems include metagenomes (red) and the biosphere-level network (maroon). Fits for each dataset (solid lines) are shown with 95% confidence intervals (dashed lines). For reference, shown in light grey is data for all biochemical networks (individuals, ecosystems, biosphere). Additional measures shown in Fig. S6.

Property	Distinguishable Levels of Organization (p- value)	Domain Prediction Accuracy (when normalized to network size)
Number of Reactions, N_R	Yes (10^{-6})	0.908
Number of Enzyme classes, N_{EC}	Yes (10^{-6})	0.900
Average Betweenness (nodes), $<B_{Nodes}>$	No (0.272)	0.918
Average Betweenness (edges), $<B_{Edges}>$	No (0.185)	0.924
Number of Edges (LCC), N_{Edges}	Yes (10^{-6})	0.897
Mean Degree (LCC), <k>	Yes (10^{-5})	0.908
Mean Clustering Coefficient (LCC), <C>	Yes (0.00853)	0.690
Average Shortest Path Length (LCC), <l>	No (0.26893)	0.921
Assortativity (LCC) <r>	No (0.0761)	0.635
Attribute Assortativity (LCC), $<r_{Attribute}>$	No (0.0563)	0.332
Number of Nodes (LCC), $N_{Compounds}$	NA	0.601

Table 1: Distinguishability of levels of biological organization based on scaling of global topological organization and the predictive power of global topological measures. Scaling of several topological measures distinguishes the ecosystem level networks from those of individuals (middle column). Many topological features describing individuals, once normalized by network size, are predictive of domain, where predictive accuracy is determined using multinomial regression (right column) (see Supplement methods). Network size in itself is insufficient to predict domain.

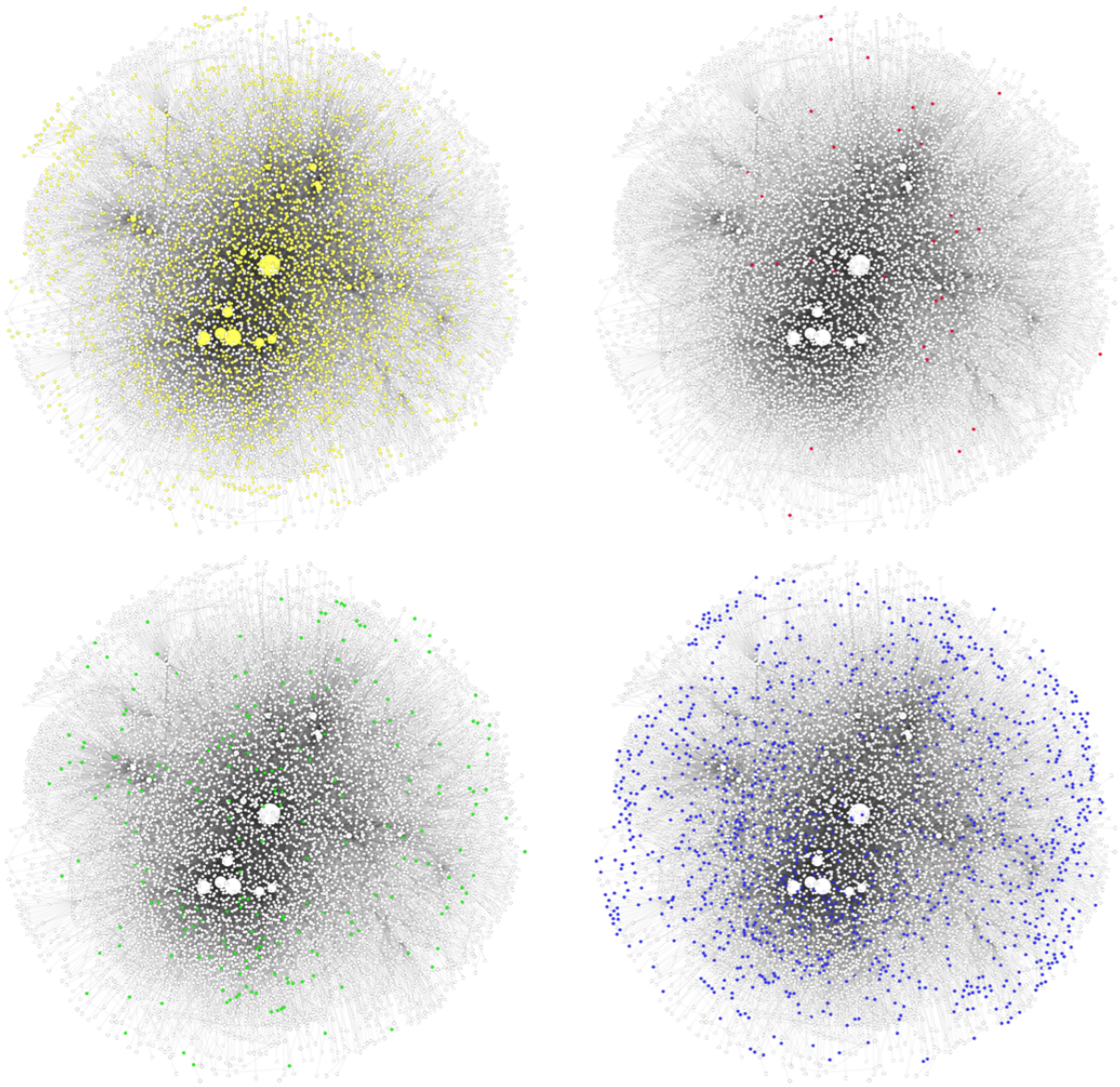

Fig. 2: Networks for three domains of life embedded within the biosphere level network. Each panel shows the same biosphere network where nodes are white and edges are grey, highlighting compounds shared by all three domain of life (37.23%, yellow), those unique to archaea (0.44%, red), bacteria (3.14%, green) and eukarya (17.08%, blue), respectively, as determined from annotated data in the 22,559 genomes in our dataset. The size of a node represents its relative degree in the biosphere network (larger nodes have higher degree).

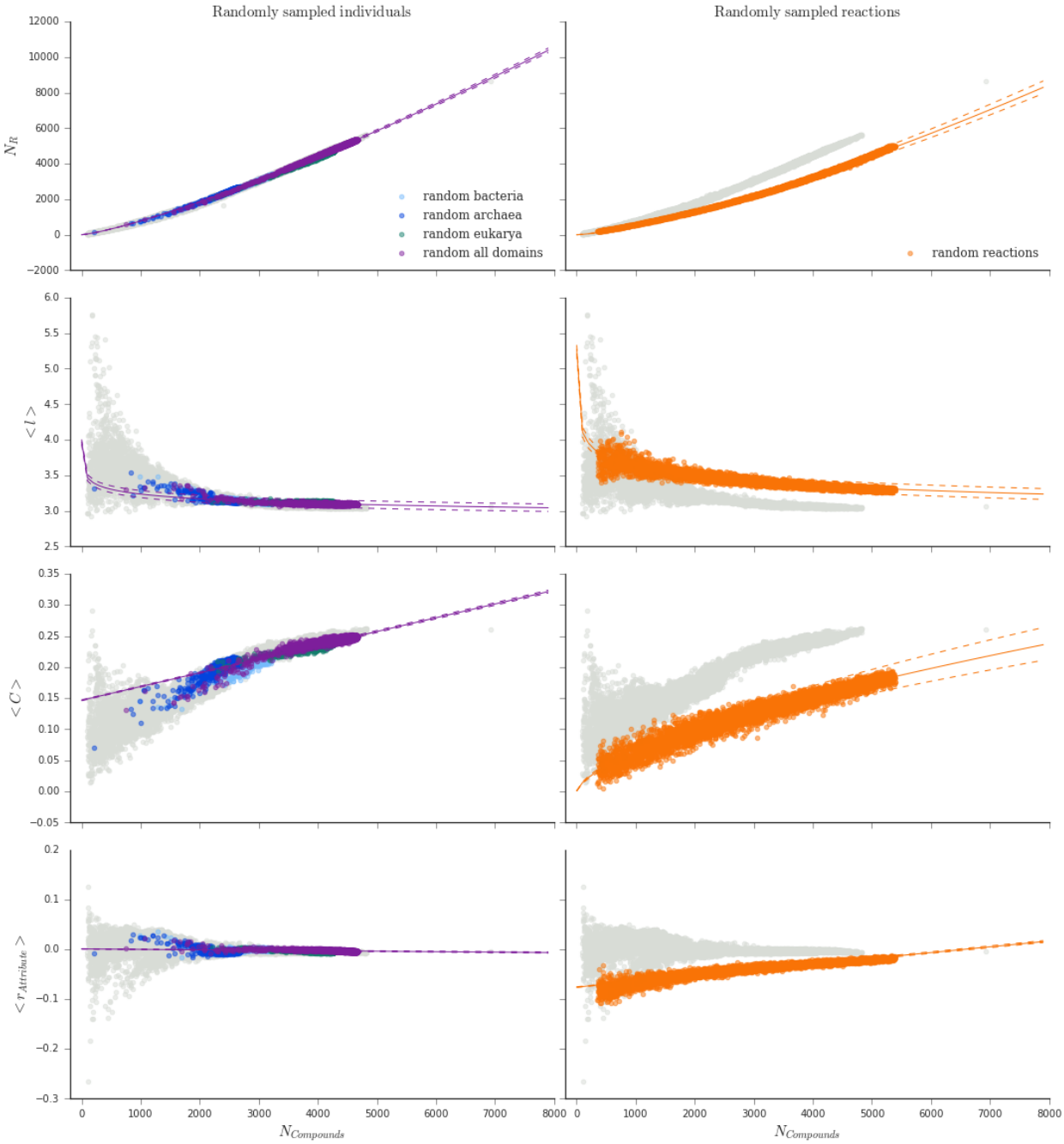

Fig. 3: Topological scaling for merged networks composed of randomly sampled individuals (blue hues, left column) or randomly sampled reactions (orange, right column). Measure and fit descriptions match those described in Fig. 1. Merged networks composed of individuals include bacteria only (light blue), archaea only (dark blue), eukarya only (blue-green), and all domains combined (purple). Merged networks composed of reactions (orange) are generated by randomly sampling reactions catalogued in KEGG. For reference, all real biological networks from Fig. 1 are shown in light grey. Additional measures shown in Fig. S7.