Scaling laws in enzyme function reveal a new kind of biochemical universality

Dylan C. Gagler[a], Bradley Karas[a], Christopher P. Kempes[b], John Malloy[a], Veronica Mierzejewski[a], Aaron D. Goldman[c], Hyunju Kim[a,d,e,1], and Sara I. Walker[a,b,d,e,1]

[a]School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85281; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]Department of Biology, Oberlin College, Oberlin, OH 44074; [d]Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ 85281; and [e]ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, AZ 85281

Edited by Eugene Koonin, National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD; received April 7, 2021; accepted December 18, 2021

All life on Earth is unified by its use of a shared set of component chemical compounds and reactions, providing a detailed model for universal biochemistry. However, this notion of universality is specific to known biochemistry and does not allow quantitative predictions about examples not yet observed. Here, we introduce a more generalizable concept of biochemical universality that is more akin to the kind of universality found in physics. Using annotated genomic datasets including an ensemble of 11,955 metagenomes, 1,282 archaea, 11,759 bacteria, and 200 eukaryotic taxa, we show how enzyme functions form universality classes with common scaling behavior in their relative abundances across the datasets. We verify that these scaling laws are not explained by the presence of compounds, reactions, and enzyme functions shared across known examples of life. We demonstrate how these scaling laws can be used as a tool for inferring properties of ancient life by comparing their predictions with a consensus model for the last universal common ancestor (LUCA). We also illustrate how network analyses shed light on the functional principles underlying the observed scaling behaviors. Together, our results establish the existence of a new kind of biochemical universality, independent of the details of life on Earth's component chemistry, with implications for guiding our search for missing biochemical diversity on Earth or for biochemistries that might deviate from the exact chemical makeup of life as we know it, such as at the origins of life, in alien environments, or in the design of synthetic life.

scaling laws | biochemical networks | astrobiology | statistical physics | enzymes

Life emerges from the interplay of hundreds of chemical compounds interconverted in complex reaction networks. Some of these compounds and reactions are found across all characterized organisms (1), informing concepts of universal biochemistry and allowing rooting of phylogenetic relationships in the properties of a last universal common ancestor (LUCA) (2). Thus, universality as we have come to understand it in biochemistry is a direct result of the observation that all known examples of life share common details in their component compounds and reactions. However, this concept of universality is quite different from universality in other fields of research, such as in the physical sciences. For example, in statistical physics, universality describes properties or macroscopic features observed across large classes of systems irrespective of the specific details of any one system (3). Universality classes become apparent in certain limits where common patterns emerge in the statistics of large numbers of interacting component parts. In some cases, the identified universality classes can be characterized by common exponents in the power laws relating different features of a given system. When a universality class is identified with distinct exponents governing its scaling behavior, the discovery can allow its predictions to guide the search for new examples (e.g., in materials discovery). Correspondingly, if biochemistries could be shown to be representative of universality classes in the physical sense, a mechanistic understanding of the identified scaling exponents could have important implications for informing models of new examples of life, beyond the specific biochemistry and evolutionary history of life as we know it.

It is an open question whether features of biochemistry can be abstracted to demonstrate behavior consistent with characterization into a universality class (or classes), but there is good reason to suspect this might be possible. Physiology across diverse organisms is already known to follow power law scaling relationships (4). These are often explained by evolutionary minimization of the costs associated with hard physical limits, such as those set by the laws of diffusion, gravitation, hydrodynamics, or heat dissipation (5). Furthermore, biochemical systems are known to display universal structure across the three phylogenetic domains in the reported scale-free (power law) connectivity of compounds within biochemical reaction networks (6). It has also been shown that scaling laws apply across networks; the average topological properties of biochemical networks follow scaling relations across examples drawn from individuals and communities (7). It is, therefore, reasonable to conjecture that the evolution of the biochemical components, which compose these networks, could be subject to physical constraints that would exhibit telltale scaling relationships indicative of universal physical limits on their collective properties.

Significance

Known examples of life all share the same core biochemistry going back to the last universal common ancestor (LUCA), but whether this feature is universal to other examples, including at the origin of life or alien life, is unknown. We show how a physics-inspired statistical approach identifies universal scaling laws across biochemical reactions that are not defined by common chemical components but instead, as macroscale patterns in the reaction functions used by life. The identified scaling relations can be used to predict statistical features of LUCA, and network analyses reveal some of the functional principles that underlie them. They are, therefore, prime candidates for developing new theory on the "laws of life" that might apply to all possible biochemistries.

Author contributions: D.C.G., C.P.K., H.K., and S.I.W. designed research; D.C.G., B.K., J.M., V.M., A.D.G., and H.K. performed research; A.D.G. contributed new reagents/analytic tools; D.C.G., B.K., C.P.K., A.D.G., H.K., and S.I.W. analyzed data; and D.C.G., C.P.K., A.D.G., H.K., and S.I.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: hkim78@asu.edu or sara.i.walker@asu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2106655119/-/DCSupplemental.

Published February 25, 2022.

Enzyme functions are a good first candidate to look at to determine if biochemistry exhibits universality in a scaling limit. Enzymes play a central role in biochemistry, catalyzing the majority of cataloged biochemical reactions. While there exist different classification systems for characterizing enzyme functions, the most thoroughly developed and widely adopted is the Enzyme Commission Classification scheme. Enzyme Commission numbers organize enzyme-catalyzed reactions hierarchically using four-digit numerical classifiers, which systematically categorize enzyme functions by their reaction chemistry (*SI Appendix*, Table S1). Take for example, identifier 1.1.1.1, which is the four-digit identifier for the alcohol dehydrogenase reaction. The identifier 1.x.x.x labels the class of the enzyme as oxidoreductase, 1.1.x.x specifies the subclass of oxidoreductases using CH-OH groups as electron donors, 1.1.1.x specifies the sub-subclass using CH-OH groups as electron donors with NAD+ or NADP+ as electron acceptors, and 1.1.1.1 is the specific Enzyme Commission number when an alcohol is the substrate (e.g., the alcohol dehydrogenase reaction). Each enzyme with known function is assigned an Enzyme Commission number for its function(s). In this way, the enzyme classification scheme provides a codified binning, or "coarse graining" in physics terminology, of biochemical reaction space, where the specification of each additional digit in the Enzyme Commission number refers to an increasingly fine-grained specification (smaller bin size) of enzyme-catalyzed reactions with common functional features.

In physics, the notion of coarse graining is critical to identifying universality classes because it allows one to disregard most details of individual systems in favor of uncovering systematic behavior across different systems. At the coarsest scale of the first digit, which corresponds to the enzyme class (EC), most details specific to individual reactions are ignored. For example, in the case of oxidoreductases, the details of the donor and acceptor do not matter—the only detail relevant to classification as an oxidoreductase is that the reaction involves electron transfer. Biochemical reactions are grouped into seven ECs as designated by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (8). Despite this codification of biochemical reaction space and a natural interpretation in terms of the coarse graining of catalytic function in biochemistry, there have been relatively few analyses to determine whether or not systematic trends in ECs exist across biochemical systems. However, if universality can be shown to apply across a large cross-section of biological diversity with respect EC functions, it would provide the strongest candidate yet for defining biochemical universality classes, akin to universality classes in the physical sense and allowing for the prediction of properties of unobserved or to be engineered biochemistries. In this manuscript, we demonstrate exactly this kind of universal behavior, which we expect arises due to optimization against hard physical limits that should similarly constrain other examples of life in the universe, including application to synthetically designed life.

Results

Scaling Laws in Enzyme Function Define Universal Properties across Diverse Biochemical Systems. We acquired genomic and metagenomic data from the Department of Energy Joint Genome Institute's Integrated Microbial Genomes and Microbiomes (DOE-JGI IMG/M) database (9, 10). Methods for filtering the dataset to remove under- or overannotated samples are described in *Data Filtering*. Our filtered data include 11,955 metagenomes, 1,282 archaea taxa, 11,759 bacteria taxa, and 200 eukaryotic taxa as well as 5,477 enzyme functions cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which we use as a proxy for the estimated aggregate number of each EC in the biosphere based on known functions. By studying ensembles of biochemical systems at the level of individuals (genomes), ecosystems (metagenomes), and planetary wide (all functions in KEGG), we are better positioned to identify universality classes that are scale invariant. We are interested in scale invariance because any scaling laws identified will be much more likely to apply to new examples of life if they are first shown to apply across all known examples, independent of the scale at which we study them.

Each of the primary classes defined by the NC-IUBMB (8) specifies a major group of enzymatic reactions. These include EC 1 oxidoreductases, EC 2 transferases, EC 3 hydrolases, EC 4 lyases, EC 5 isomerases, EC 6 ligases, and EC7 translocases (Fig. 1 and *SI Appendix*, Table S2). In what follows, we consider only ECs 1 through 6, as EC 7 (translocases) was only recently added, and at the time of writing, it is insufficiently annotated to allow for rigorous statistical analysis across the data available to this study. We analyzed scaling patterns at the macroscopic scale of ECs by counting the number of unique Enzyme Commission identifiers within a given EC (as designated by the first digit) across biochemical datasets representing different phylogenetic domains and levels of organization (Fig. 1). By focusing on ECs, our analysis considers only the grouped, or coarse-grained, functionality of biochemical reactions, ignorant of their more detailed mechanisms. This coarse graining of biochemical reaction space is analogous to coarse graining in physical systems, where macroscale observables, like temperature, have been shown to be more effective (predictive) descriptions than considering the multitude of possible microstates, permitting the derivation of quantitative behavior that can predict behavior across many systems.

We plotted the total number of unique enzyme identifiers within a given EC ("EC numbers in EC class") as a function of the total number of unique identifiers across all classes ("total EC numbers") for each annotated genome or metagenome and for the biosphere (all KEGG reactions). The resultant empirically determined scaling behaviors are shown in Fig. 2, where each data point represents the binned statistics of ECs for a given genome, metagenome, or the biosphere. These scaling relationships capture systematic changes in the number of functions within a given EC, relative to the total number of unique functions across all enzymes in an organism or ecosystem. We observe regular scaling behaviors for each EC across biochemical systems as they increase in the size of their reaction space. Both linear regression models and power law models were fit to the data; the power law models consistently were either equivalent to or outperformed the linear regression models using an SE minimization test (*Fitting Scaling Laws to Empirical Data*).

We find that all ECs display scaling behavior with positive exponents ($k > 0$) with the power law fit, $y = ax^k$, such that the total number of unique functions in each EC systematically increases with an expanding total number of enzyme functions, with some ECs increasing much more slowly than others. Before describing these trends, it should be emphasized that many biological features do not follow scaling relationships. The observation of scaling itself is, therefore, not trivial and indicates a certain type of organizing mechanism (5). For example, genome size in mammals and other metazoan classes does not strongly scale with body size, and unit repair costs are roughly invariant across all of life (11, 12). Similarly, some aspects of biochemistry that are largely conserved across all of life, such as the genetic code or translational machinery, also clearly do not exhibit scaling relationships in the diversity of their structure.

We classify the observed scaling behaviors by their scaling coefficient (with associated CI) into three categories: sublinear, $k < 1.0$; linear, $k = 1.0$; and superlinear, $k > 1.0$. Scaling behavior consistent with a linear fit ($k = 1.0$) indicates a fixed ratio of functions within a given functional group to total enzymatic functions, whereas sub- or superlinear behavior is indicative of a depletion or enrichment of functions within a given EC, respectively. Based on the sublinear, linear, and superlinear

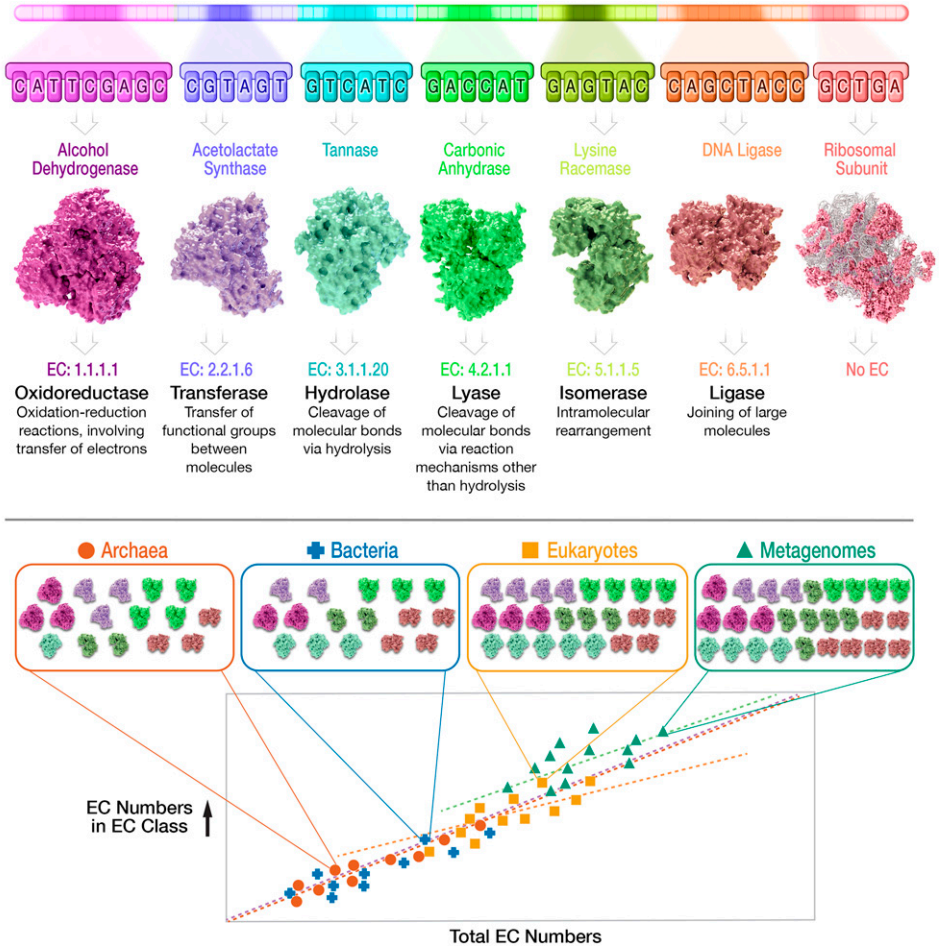

Fig. 1. Conceptual schematic showing how genomic and metagenomic data are used to determine bulk trends in the number of enzyme functions for each major ECs. In data archived by the JGI, many genes in genomes or metagenomes have been identified, and functions assigned to protein-coding regions are mapped to specific four-digit Enzyme Commission number identifiers. For each genomic or metagenomic sample, we then binned Enzyme Commission numbers based on the primary digit in the identifier, which specifies the EC as an oxidoreductase (EC 1), transferase (EC 2), hydrolase (EC 3), lyase (EC 4), isomerase (EC 5), or ligase (EC 6). Scaling relations are then determined by counting the total number of unique enzyme functions within a given EC (EC count) as a function of total enzyme functions across all ECs (total enzymes [abbreviated for total enzyme functions]). Results are compared across the three domains archaea, bacteria, and eukaryota and across metagenomes.

classification, we can determine whether specific ECs display universal behavior across the different datasets (Table 1). If each dataset shares the same classification for a given EC across all datasets, then that grouping of functionality is a good candidate for a universality class. We are motivated to identify such classes as the strongest candidates for biochemical universality because the divide between sublinear and superlinear scaling is an important one; it can represent distinct mechanisms or optimization for distinctly different types of constraints. ECs with the same scaling classification across all datasets, therefore, make possible a unified description in terms of the same underlying principles across all known examples of biochemistry. This strong case is meant to delineate examples that make closest contact with universality classes as we know them in physics, where observations of the same properties across different systems are described by similar underlying mechanisms. Conversely, an empirically observed scaling behavior is a weak candidate for a universality class if the scaling law does not share the same classification across datasets, such that we might not expect these to be readily describable by universal mechanisms.

Two ECs, the lyases (EC 4) and isomerases (EC 5), change their classification moving from prokaryotes to eukaryotes (from superlinear to linear for lyases and from sublinear to linear in isomerases). For the isomerases, which are sublinear in all other categories, candidacy as a universality class cannot be ruled out because the CIs on the scaling exponent in eukaryotes also include a sublinear fit. The lyases, which are consistent with superlinear scaling across the domains, are sublinear for metagenomes with a CI that does not overlap that of a superlinear fit. Therefore, it is only for the lyases that we see weak evidence for a universality class based on the empirical data. This could be attributable to the relatively low diversity in unique functions among the lyases as compared with other EC classes (*SI Appendix*, Table S2). However, the ligases are even less diverse, yet for this class, we do see universal behavior (isomerases also show less diversity, and universality cannot be ruled out). In total, five of the six classes exhibit strong evidence for universality within CIs for the scaling law fits. These are the oxidoreductases (EC 1) and hydrolases (EC 3) both exhibiting superlinear scaling across all datasets, transferases (EC 2) and ligases (EC 6) exhibiting sublinear scaling behavior across all data, and the isomerases (EC 5) for which sublinear scaling across all data cannot be ruled out (Table 1). The results are consistent with the possibility that cataloged biochemistry is part of a universality class in each of these five ECs.

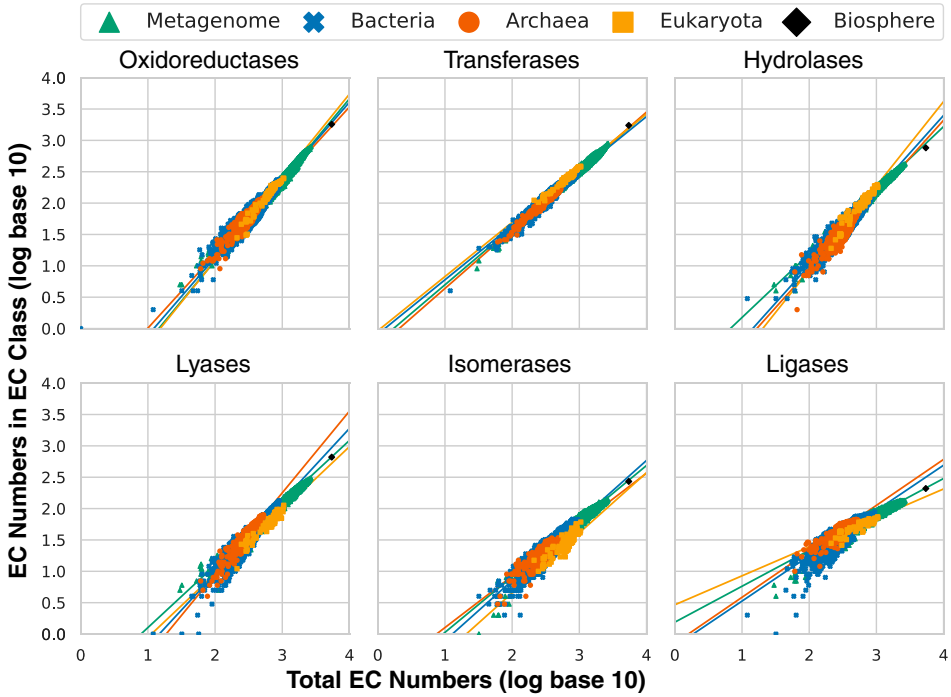

Fig. 2. Scaling behaviors in enzyme function as captured by the number of enzyme functions within each major EC (EC numbers in EC class) as a function of total enzyme functions (total EC numbers). Shown are the ensemble statistics for biochemical data derived from annotated genomes sampled from archaea, bacteria, and eukaryota taxa and from annotated metagenomes. The oxidoreductases and hydrolases display superlinear scaling across all four biochemical ensembles, whereas for transferases and ligases, sublinear scaling is observed universally (Table 1).

Within a dataset, we generally observe small 95% CIs on the scaling exponents (*SI Appendix*, Tables S4–S6 show fit data, including fits to all three domains [pantaxa] and all data). For several ECs, the scaling exponents are significantly different across datasets, particularly between prokaryotes and eukaryotes. For example, the exponent for ligases, while consistently sublinear, changes by nearly 1/4 between prokaryotes and eukaryotes, which could have significant implications for differences in the diversity of polymerization reactions across different biochemical systems. The most consistent exponent values across taxa and metagenomes are for the transferases and isomerases. Lyases have the largest variation in coefficients and the largest shifts in classification, consistent with their weak evidence of universality in their scaling behavior. Why only the lyases stand out in this regard is a subject of interest. With respect to our inspiration from statistical physics, it could be that lyases as a group are not a natural partition of reaction space with respect to the constraints that operate on biochemical organization (e.g., in thermodynamics, temperature and pressure are appropriate variables, while other possible measures are not mathematically well behaved). To test this hypothesis, we grouped the lyases together with hydrolases because both exhibit similar function; these two classes describe reactions that break down molecules, with the hydrolases a special subset that involves water. We find that the combination of lyases and hydrolases together exhibits universal superlinear scaling across datasets, although metagenomes do exhibit a slope close to linear (*SI Appendix*, Table S8). Thus, by grouping

Table 1. Regression values of the power law fits for the slope shown in Fig. 2 as 95% CIs

	Archaea	Bacteria	Eukaryota	Metagenome
Oxidoreductases	[1.152, 1.199]	[1.233, 1.245]	[1.290, 1.364]	[1.286, 1.295]
Transferases	[0.924, 0.951]	[0.864, 0.871]	[0.843, 0.886]	[0.908, 0.914]
Hydrolases	[1.162, 1.228]	[1.191, 1.202]	[1.299, 1.390]	[1.012, 1.017]
Lyases	[1.281, 1.325]	[1.153, 1.163]	[0.968, 1.060]	[0.992, 0.998]
Isomerases	[0.792, 0.849]	[0.952, 0.964]	[0.892, 1.025]	[0.883, 0.891]
Ligases	[0.712, 0.753]	[0.716, 0.728]	[0.430, 0.493]	[0.570, 0.576]

The data were fit to the power law relation $y = ax^k$, where k is the slope and a is the intercept of the logarithmically transformed regression. Dark shaded cells indicate fits that exhibit superlinear scaling (slope > 1.0), light shaded cells indicate linear scaling (slope = 1.0), and white cells indicate sublinear scaling (slope < 1.0).

lyases and hydrolases together, all functions coarse grain into five classes that are consistent with universal scaling behavior within CIs for the scaling fits.

These scaling relationships appear to be more universal than previously identified scaling laws for other biological observables, which often change classification across domains or across levels of organization. The scaling relationships for whole-organism metabolic and growth rates are known to dramatically shift across the bacteria/eukaryote divide (13, 14). Metabolic rates increase superlinearly with organism size for bacteria but only linearly for unicellular eukaryotes and sublinearly for multicellular eukaryotes (14). Even more strikingly, the growth rates derived from metabolic scaling increase with cell size in bacteria but decrease with cell or body size in eukaryotes (13). Similarly, there are known asymptotic limits at the large end of bacteria related to increasing the number of ribosomes to keep pace with growth rates, which implies another dramatic shift across this evolutionary transition (15). Despite these differences in other physiological observables, we find that ECs follow relatively consistent scaling behavior both across taxa and across the level of ecological organization. Of note, the relative ordering of increasing or decreasing values for the scaling exponents follows roughly the ordering of divergence times for the three domains (*SI Appendix*, Table S9).

The enzyme function scaling laws for single organisms hold promise for predicting missing biochemical diversity in the form of missing enzyme functions at the metagenome or biosphere levels of organization and potentially, for underannotated organisms. For example, comparing projected trends in Fig. 2 with the currently cataloged number of commission numbers in each class indicates that there are many oxidoreductase and hydrolase functions left to be discovered, whereas the scaling laws predict that essentially all ligase and isomerase functions are already identified (*SI Appendix*, Fig. S6). The scaling relations underestimate biosphere-level enzyme diversity only in the transferases, possibly due to the lower representation of eukaryota and archaea in our dataset; scaling trends for these domains most closely approach the total cataloged transferase diversity. Some of the differences in the scaling exponents observed across taxa are significant and could illustrate meaningful physiological differences across evolutionary transitions. As an illustration of this, the bacterial scaling exponent would overpredict the change in the number of eukaryote lyases by a factor of 1.5 for an order of magnitude increase in the total ECs.

Universality in Scaling of Enzyme Function Is Not Explained by Universally Shared Components. A major challenge for any claim of universality observed across life on Earth, which seeks to inform more general principles, is the shared ancestry of known life. Evolutionary contingency influences the biochemical concept of universality because of a shared component set of enzyme functions, reactions, and compounds common to all life—the product of shared evolutionary history. This allows for rooting of phylogenetic relationships in the properties of an LUCA, which is expected to share much of the same universal component set. In this sense, the biochemical concept of universality pertains to what physicists refer to as microscale features, which here manifest as the specific molecules and reactions used by all life. By contrast, the scaling behaviors we have identified in the previous section pertain to a macroscale feature arising in the statistics of many biochemical reactions. The scaling relations, therefore, need not, in principle, rely on the presence of a shared component chemistry across systems. In fact, in order to identify EC scaling as a universality class in the physics sense, which could allow us to generalize beyond life as we know it, a key requirement is that universality in scaling behavior does not directly depend on universally shared component chemistry. We, therefore, sought to determine whether the presence of universal scaling laws is strongly driven by the presence of universal biochemical components. We include LUCA as a model of early life, which is itself constructed based on the existence of shared components across modern systems (16). The consensus LUCA model we use for comparison is derived from the eight current leading models of LUCA (*Materials and Methods*).

Our first goal was to determine if the scaling laws in Fig. 2 arise because of a set of highly redundant enzyme functions across samples in the biochemical ensembles or are instead macroscale features that do not depend on the exact functions used. If the latter is true, then they can be attributed as a convergent macroscale feature that does not depend on shared component parts. To determine this, we evaluated the component universality of each enzyme function by rank-ordering enzyme functions by their frequency of occurrence across a given dataset. The results are shown in Fig. 3. We then assigned area under the curve (AUC) scores (*Materials and Methods*) to the occurrence frequencies of enzyme functions across domains and metagenomes. These AUC values allow for efficient comparison of the distribution of unique enzyme functions within a given EC for each dataset, where values closer to AUC = 1 indicate that an EC has component enzyme functions that are more commonly distributed and values closer to AUC = 0 indicate cases where specific enzyme functions are relatively rare.

AUC results for data in Fig. 3 are in Table 2; "pantaxa" and "all" categories include data from all three domains or from all domains and metagenomes grouped together, respectively. From the frequency of occurrence curves (Fig. 3) and their AUC scores (Table 2), it is apparent that macroscale universality in EC scaling does not directly correlate with a high degree of microscale universality in enzyme function. This result can be seen most clearly by comparing Table 2 with Table 1. The oxidoreductases consistently exhibit the lowest AUC scores among ECs, corresponding to a high degree of unique enzyme functions across different biochemical systems. Yet, as a class, the oxidoreductases also exhibit universal scaling with very tightly constrained coefficients (only transferases have tighter constraints). In contrast, ligases are the most universal in terms of their components but also, have the largest variation in terms of scaling coefficients among any of the ECs. Thus, the oxidoreductases and ligases give us two end-member cases. In the first, tightly constrained universal scaling behavior emerges from ensembles with relatively few universally shared component enzyme functions (e.g., the oxidoreductases). In the second, loosely constrained universal scaling behavior emerges from ensembles with a high degree of shared component functions across samples (e.g., the ligases). These end-member cases highlight how the observed scaling trends cannot be explained directly by the universality of the underlying component functions. In fact, looking across all classes, there is no direct correlation between universality of unique functions within a given EC and universality in the apparent scaling behavior of the EC (*SI Appendix*, Fig. S15). This lack of correlation indicates that EC scaling is indeed a macroscale property that emerges in the statistics of many reactions and therefore, could reflect universal physical constraints on the architecture of biochemical networks in terms of their catalytic functional diversity. Metagenomes have the highest AUC, meaning that it is more likely to find common functions across community samples than across individuals sampled from the three domains. However, we also observe that metagenomes share similar EC scaling to individuals for some ECs (a key part of our interpretation of universality classes), further corroborating a lack of direct correlation between scaling patterns and AUC scores.

We also sought to understand the relationship between the EC diversity in a consensus model of LUCA and the

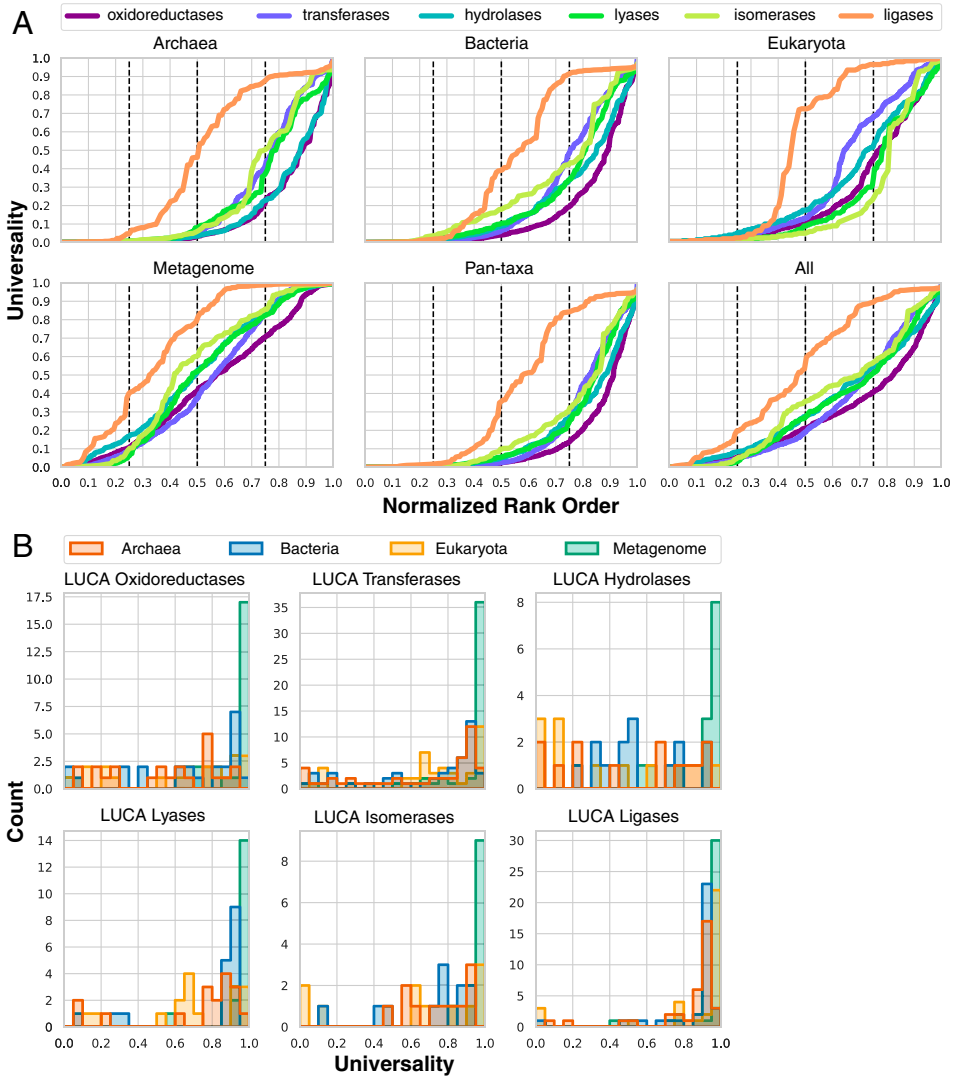

Fig. 3. (*A*) Universality of specific Enzyme Commission number identifiers for each EC. Unique enzyme functions are rank ordered according to their frequency of occurrence across a given dataset. (*B*) Distribution of the 154 Enzyme Commission numbers present in the consensus LUCA model as found across each of our datasets. The LUCA is more universally distributed across metagenomes than individuals, and the overall shape of the distributions, including skew toward being more universally distributed, depends on the EC.

universality class of modern biochemistry as dictated by our identified EC scaling relations. We compared the distributions of LUCA enzyme functions with their universality across modern biochemical systems (Fig. 3*B*). We find that the universality of LUCA enzyme functions in modern systems depends strongly on the EC class and the dataset. Across all six ECs, most component enzyme functions found in LUCA are also found in modern metagenomes, corroborating proposals that inferred LUCA genomes represent a population of organisms rather than an individual (17). Comparing the three domains, the patterns vary significantly by EC. For superlinear classes—the oxidoreductases and hydrolases—functions in LUCA are nearly uniform in terms of their distribution across modern organisms; some functions are rare, and some are common. The other four ECs exhibit a bias toward functions from LUCA that are more universal across modern organisms. This bias is most evident in the case of the ligases, where nearly all ligase functions in the LUCA model are found in >90% of modern organisms across all three domains. This is to be expected based on how LUCA models are currently constructed; since LUCA is phylogenetically reconstructed or at least reconstructed from consensus across extant organisms, the commonality of ligase functions across different biochemical systems means they are more likely to be represented in LUCA models than other ECs, which have fewer universal functions.

We next compared the distribution of total enzyme functions, reactions, and compounds found in LUCA in terms of their universality across modern life to determine what specific components are more or less common across our datasets and to corroborate what specific dataset(s) are more representative of the consensus LUCA (*SI Appendix*, Fig. S14). Independent of domain or level of organization, biochemical systems tend to share more compounds in common than they do reactions and more reactions than enzyme functions ($AUC_{compounds} > AUC_{reactions} > AUC_{enzyme\ functions}$ across all datasets) (*SI Appendix*, Fig. S14, *Upper* and Table S10). Of the 154 enzymes, 337 reactions, and 438 compounds in the consensus LUCA, there is (as should be expected based on LUCA's phylogenetic construction) a strong bias in the distribution toward inclusion of more universally distributed components. As with EC distributions, the majority (>90%) of compounds, reactions, and enzyme functions in LUCA are found in >90% of metagenomes. Again, we can conclude that LUCA components are more universally found in metagenomes than they are in any of

Table 2. AUC scores for data shown in Fig. 3

Domain	Oxidoreductases	Transferases	Hydrolases	Lyases	Isomerases	Ligases
Archaea	0.152	0.244	0.158	0.228	0.248	0.461
Bacteria	0.156	0.249	0.210	0.233	0.284	0.431
Eukaryota	0.253	0.337	0.303	0.233	0.214	0.522
Metagenome	0.431	0.451	0.508	0.479	0.518	0.663
Pan-taxa	0.129	0.190	0.173	0.189	0.214	0.405
All	0.270	0.311	0.330	0.320	0.357	0.526

A score of AUC = 1 means all enzyme functions occur in 100% of samples and a value of 0 indicates functions are found in none. Thus, AUC scores closer to 1 indicate more universality in the distribution of specific functions within a given class. Shading indicates superlinear, linear or sublinear scaling behavior observed for a given class across a given data set, as in Table 1.

the three domains, consistent with hypotheses that LUCA should best be understood as an ecosystem-scale property (17–19).

The next step was to compare the EC distribution in LUCA with that of modern life. Fig. 4 shows where the consensus model of LUCA falls with respect to projected fits of EC scaling down to the size scale of 154 enzyme functions in the consensus model. The CIs by dataset and EC are in Table 3. The only dataset to yield accurate predictions across all ECs (within 95% CIs) is archaea. This could arise as a selection bias as archaea are less well characterized than other domains and therefore, include fewer lineage-specific elaborations. However, the capacity of scaling laws to reconstruct a potential macroscale similarity between archaeal functions and LUCA functions also appears to be independent of common component chemistry; the ensemble dataset of archaea shares fewer universal components than the other domains, metagenomes, or pantax datasets, indicating that the predictive capacity of archaeal scaling laws is reconstructing bulk properties of phylogenetic relationships without requiring a high degree of shared component parts. By contrast, we find that metagenomes are the least predictive of the datasets, with LUCA's functional diversity outside of the projected values for every EC. This lack of predictive accuracy is somewhat surprising given that all LUCA functions are found in almost all metagenomes (Fig. 3) but is perhaps less surprising when one considers that current efforts to reconstruct LUCA often focus on our universal ancestor as an individual-scale biochemistry rather than a scale-invariant one. Future directions could include using the scaling laws identified here to constrain properties of ensemble models of LUCA that exist at the ecosystem scale. Overall, our results indicate that universality in LUCA components in modern life does not immediately imply that LUCA shares the same universality class for those components, corroborating our conclusion that enzyme scaling is not tied strictly to the presence of common components and therefore, can provide a useful tool for better constraining properties of the earliest life on Earth.

Network Analyses Suggest Functional Principles Underlying Enzyme Scaling Behaviors. A key consideration in identifying new scaling laws is to determine the functional principles underlying them. If the scaling laws reported herein are indeed universal to all biochemical systems, understanding the functional principles could provide a critical tool for predicting the properties of life not yet observed. To take steps in this direction, we studied the statistical properties of ECs as related to biochemical network topology.

If one feature of a system has a larger scaling exponent than another, this often implies that as systems' size increases, this feature will grow more quickly. This must be the case if the cross-system scaling relationships are preserved in time (20). Thus, we can interpret underlying mechanisms of the observed scaling relationships by considering network growth dynamics. A priori, we would expect the oxidoreductases to expand most rapidly in growing networks and the ligases to expand least rapidly to be consistent with the interspecific scaling that we have demonstrated. We performed network expansion (21, 22) on the biosphere-level network and tracked EC diversity as the network expanded, and this is indeed what we find. In network expansion, one starts from a "seed set" of handfuls of compounds and reacts them; the products produced are then added to the list of possible reactants, and the algorithm is iteratively repeated until no new products are formed. It has been used to study early life (23, 24) and the evolution of Earth's biochemistry over geological timescales (22), as well as on other planetary bodies (25).

We performed network expansion and grouped enzymes by their EC. Early growth in the expansion of all six ECs leads to separation of classes later in the expansion; sublinear ECs tend to contribute to growth of the network only at early time steps, leaving superlinear ECs to dominate the latter growth of the network (*SI Appendix*, Fig. S18). However, this separation of network growth patterns by class is not so clear when using the traditional EC classifications. Therefore, we next performed an experiment where we tracked lyases and hydrolases together, motivated by our results confirming that these two classes together exhibit universal superlinear scaling behavior (*SI Appendix*, Fig. S13 and Table S8). Our expectation was that enzyme groupings that exhibit universality in their scaling behavior might exhibit clearer signals in expansion on the biosphere-level network. This is indeed the case (Fig. 5), where growth of the network in the expansion process is dominated by the combined lyases and hydrolases early on and the oxidoreductases later. These two classes (lyases and hydrolases combined and oxidoreductases) are both superlinear. The sublinear classes with the smallest scaling coefficients—the isomerases and ligases—are the two classes that contribute least to driving the expansion at late times. Ligases and isomerases also exhibit the highest AUC scores across datasets (Table 2), suggesting that their conserved function might arise because they are not contributing to newer functions as the biosphere has expanded its functionality. Transferases are an intermediate case as they exhibit sublinear scaling behavior, but their growth in network expansion is still significant at late times, although not at the rates of lyases and hydrolases (early) or oxidoreductases (late). To further corroborate these results, we also performed network expansion experiments with random seed sets (*SI Appendix*, Fig. S19). The results are consistent with a general pattern where grouping into the five universal classes (oxidoreductases, lyases and hydrolases, transferases, isomerases, ligases) corresponds to growth patterns in network expansion predominantly driven by the superlinear classes and the transferases, where contributions to growth taper off roughly in order of the values of their scaling coefficients (i.e., ligases first with the smallest coefficient and oxidoreductases last with the largest).

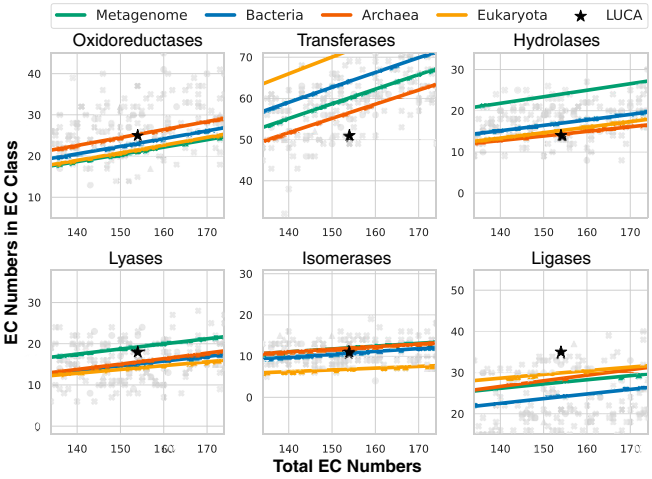

Fig. 4. Enzyme functions in a consensus model for the LUCA are consistent with the universal scaling only for some ECs and datasets. LUCA is consistent (within the 95% CI) for all projections based on EC diversity predicted by archaeal scaling laws, and metagenomes are not predictive for any EC classes (Table 3).

To further elucidate these core/periphery considerations as mechanistic explanations for the observed scaling trends, we constructed enzyme–enzyme networks where the web of biochemical reactions and compounds is projected onto the space of enzymes (*Materials and Methods*). We studied global patterns in the frequency distribution of the degree of each enzyme (e.g., the number of other enzymes connected to that node) across each dataset, which provides a window into the role of enzymes of a particular class in local connectivity (Fig. 6), where we group lyases and hydrolases together based on our earlier results (*SI Appendix*, Fig. S21 shows data on all individual ECs). The degree distributions of biochemical networks are already well known to exhibit heavy-tailed distributions. To compare different functional classes of enzymes, we fit a power law distribution to each dataset degree distribution. For the two superlinear universality classes, we observe more high-degree enzymes as compared with other classes (Fig. 6 and *SI Appendix*, Table S10). Fig. 6 shows that the slope of the degree distribution in the logarithmically transformed regression for the two superlinear classes monotonically increases as the general size of biochemical systems increases (i.e., the slope sizes are monotonically ordered by size such that $k_{Archaea} < k_{Bacteria} < k_{Eukaryota} < k_{Metagenomes}$, meaning that larger biochemical systems tend to have more high-degree nodes). We do not observe this monotonic increase in the sublinear universality class of transferases (*SI Appendix*, Table S11)—in fact, we see the opposite trend. The topological differences between sublinear transferases and superlinear oxidoreductases and hydrolases + lyases suggest that only those enzyme groups that exhibit superlinear scaling behavior tend to play a prominent role in maintaining the structural connectivity of biochemical networks as they also drive growth in size.

We also measured the betweenness centrality (26) (*Materials and Methods*) of each enzyme across datasets to capture features of influential enzymes in larger-scale connectivity patterns (*SI Appendix*, Fig. S21 has results for degree and betweenness including pantaxa data, and *SI Appendix*, Fig. S20 shows the original six Enzyme Commission Classes). Unlike the degree centrality, the distributions of betweenness centrality show little correlation with EC scaling behaviors. The superlinear oxidoreductases and the sublinear transferases show similar patterns across all datasets, while the sublinear class of ligases and superlinear class of lyases and hydrolases share similar patterns.

Taken together, our network analyses indicate that superlinear scaling laws correspond to macroscale patterns in functions that drive growth while maintaining local connectivity of biochemical networks. Functional classes that play a role only in early network growth (ligases, isomerases) play a lesser role in local connectivity. Transferases, a sublinear class, represent an intermediate case; they contribute significantly to later network expansion, but they are distinguished from the superlinear classes in that they have fewer high-degree nodes and therefore, contribute less to maintaining local network connectivity as size increases. Transferases contribute heavily to the diversity of enzyme functions in the biosphere, but the dominant contributions of new diversity tend to be led by lyases, hydrolases, and the oxidoreductases. It may be the case that our biosphere has not evolved long enough to observe the drop-off in transferases that this class's sublinear scaling behavior and topological role might indicate will occur.

Table 3. EC numbers per class for LUCA and projected numbers for each domain according to their regression values from *SI Appendix*, Table S5

	LUCA	Archaea	Bacteria	Eukaryota	Metagenome	Pan-taxa
Oxidoreductases	25	[19.6, 32.5]	**[21.5, 24.8]**	[14.1, 32.8]	**[20, 22.3]**	[22, 25.1]
Transferases	51	[48.8, 65.5]	[61.7, 66.6]	**[56.3, 91.2]**	**[58, 62.3]**	**[59.9, 64.5]**
Hydrolases	14	[10.0, 20.5]	**[15.9, 18.1]**	[9.0, 25.7]	**[23.3, 24.8]**	**[14.9, 16.9]**
Hydrolases + Lyases	32	[25.1, 37.2]	[31.3, 34.0]	[21.4, 40.5]	**[42.6, 44.5]**	[31.1, 33.6]
Lyases	18	[12.2, 19.9]	**[14.3, 16.0]**	[8.4, 23.9]	**[18.5, 20]**	**[15, 16.9]**
Isomerases	11	[8.8, 16.3]	[10.0, 11.5]	[3.2, 14.6]	**[11.4, 12.6]**	[10.1, 11.7]
Ligases	35	[22.8, 35.7]	**[22.6, 25.8]**	[20.8, 43.0]	**[26.6, 28.7]**	**[24.3, 27.6]**
Total	154	[122.2, 190.4]	[146, 162.8]	[111.8, 231.1]	**[157.8, 170.7]**	[146.1, 162.6]

Brackets show the estimated 95% confidence interval using the regression values for $y = ax^k$, where a is the intercept and k is the slope of the logarithmically transformed regression. All values were calculated using 154 as x, which is the total number of ECs in the LUCA consensus model. Highlighted in bold are projections where the LUCA model lies outside of the 95% confidence interval for the scaling coefficient in that class and dataset, with dark grey background indicating cases where the LUCA prediction underestimates the value compared to the consensus model and lighter grey indicating cases where the prediction makes an overestimate.

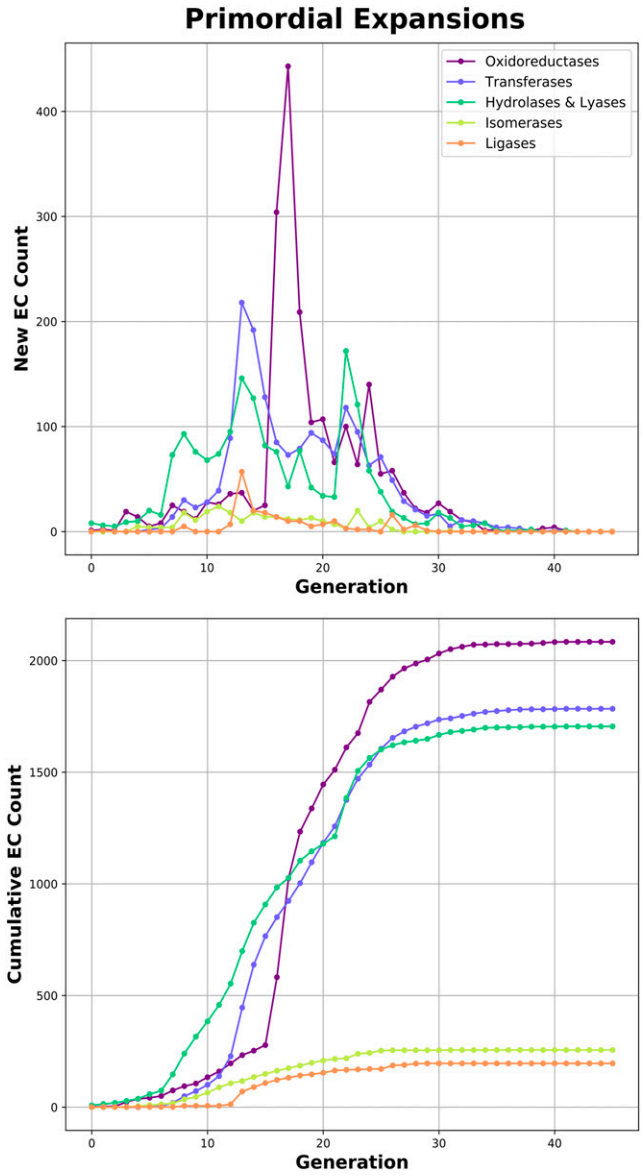

Fig. 5. Network expansion results detailing the emergence of enzyme functions at the biosphere level. (*Left*) The number of enzymes of each EC class generated per generation with a network expansion using compounds readily available in the primordial ocean (H_2O, CO_2, H_2SO_4, H_3PO_4, NH_3, and H^+) as a seed set. Hydrolases and lyases are counted together. (*Right*) The cumulative number of enzymes belonging to each EC class within the same primordial network expansion. The enzyme groupings shown taper off in their contribution to overall growth in the network roughly in order of their respective scaling coefficients.

Discussion

Traditional views on the universal nature of biochemistry have focused on the existence of specific component compounds, reactions, and enzyme functions, which are conserved across known examples of life (1). This perspective should be thought of as universality of component membership or composition. It has informed many fields of inquiry, including efforts to constrain the chemistry implicated in the origins of life and the search for life on other worlds (27). This is despite how we currently lack concrete scientific inroads to determine if the shared component chemistry of life as we know it should be universal to all life (known and unknown). However, synthetic biology is already demonstrating that perhaps component membership is not enough given experimental evidence of alternative chemistries that can function in vitro and in vivo (28, 29). In the current work, we have identified systematic regularities in the form of scaling laws that allow a different window into a new kind of biochemical universality. These are more generalizable because they do not depend on the details of components and allow for asking questions about what other biochemistries could be possible, with the constraint that they are consistent with predicted scaling trends.

Microscale details in the specific enzyme functions used by life can vary significantly from system to system; however, we have shown that the macroscale patterns that emerge from coarse-graining enzyme functions follow tightly constrained power laws. These macroscale patterns tend to correlate with features driving the global architecture and expansion of biochemical networks. This suggests a universal macroscale pattern in function across known life, which is not strictly dependent on evolutionary contingency. This universality is, therefore, likely to arise due to hard physical constraints, where the reactions used in living chemistries are universally constrained by macroscale statistics, independent of the specific catalyst (enzyme) identities. While enzymes fall into the category of biochemical macromolecules that are themselves part of the universal set of component membership, our focus on functions is not necessarily so restricted that it needs to apply solely to known biochemical catalysts. For example, many biochemical reactions have been shown to be catalyzed by alternative polymers to those used in extant life or by cofactors in origins of life studies (28, 30). Since our analyses refer only to the functions of catalysis and not the catalysts themselves, they are candidates for generalizing beyond the chemistry of life as we know it.

A critical question is whether the universality classes identified herein are a product of the shared ancestry of life. A limitation of the traditional view of biochemical universality is that universality can only be explained in terms of evolutionary contingency and shared history, which challenges our ability to generalize beyond the singular ancestry of life as we know it. Indeed, a set of closely related genomes will, by definition, share a high degree of universality in component enzyme functions. Phylogenetic effects would be a concern here too if we were claiming universality in terms of a specific set of unique enzyme functions as these then could be attributed to oversampling highly related genomes. Instead, we showed here that universality classes are not directly correlated with component universality, which is indicative that it emerges as a macroscopic regularity in the large-scale statistics of catalytic functional diversity. Furthermore, EC universality cannot simply be explained due to phylogenetic relatedness since the range of total enzyme functions spans two orders of magnitude, evidencing a wide coverage of genomic diversity. The maximum relatedness of two very different enzyme set sizes would occur in cases where the smaller set is a perfect subset of the larger. The very nature of the scaling relationships introduces diversity through set size differences.

The possibility of universal physical constraints on biochemical architecture is most apparent for ECs that are less restricted by evolutionary contingency. That is, in cases where biological systems can innovate on function (e.g., in the oxidoreductases where there is low component universality), we see tighter constraints on the empirically determined scaling behavior across domains and levels of organization. This can be contrasted with cases where evolutionary contingency plays a more significant role (e.g., in the ligases with high component universality), where we see a larger variation in observed scaling behaviors. One possible explanation is that evolutionary constraints limit optimization toward physical limits. This may indeed explain the behavior of the ligases, where optimization is likely constrained by historical

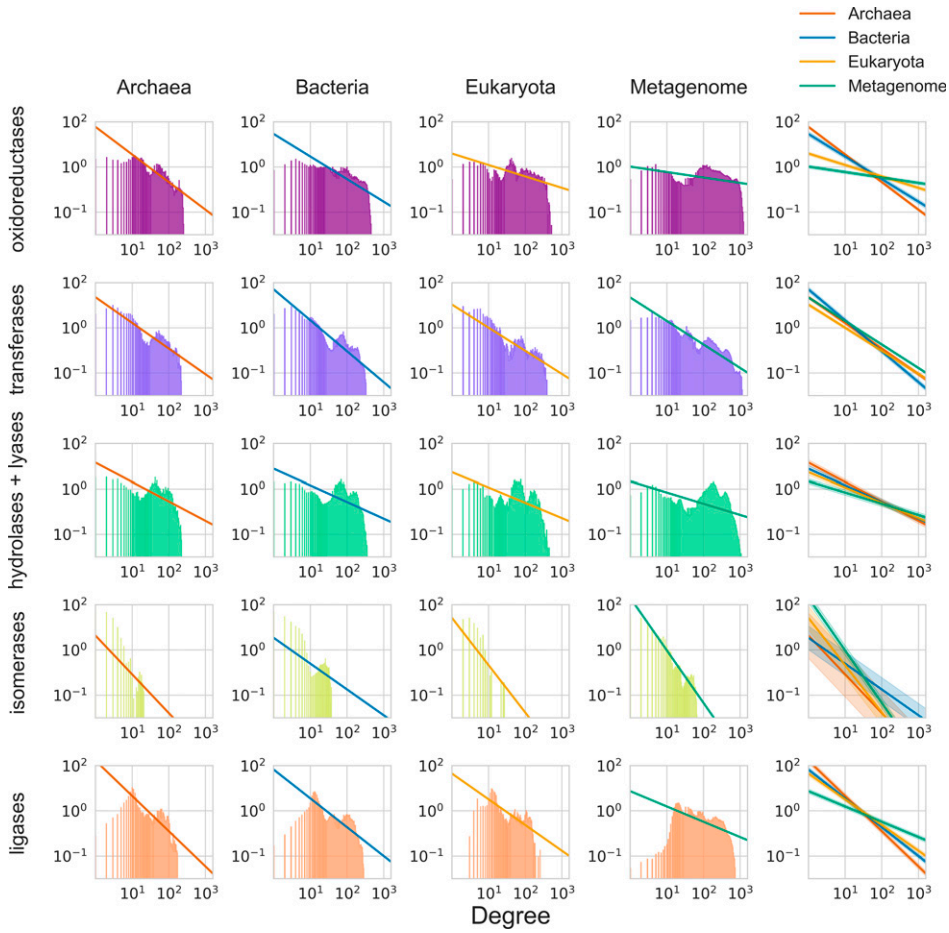

Fig. 6. Probability density distribution of degree centrality of different enzyme functions within enzyme–enzyme networks. The first four columns are associated with networks built from annotated genomes sampled from archaea, bacteria, and eukaryota taxa and from annotated metagenomes, respectively. The distribution of degrees of nodes in hydrolases and lyases is computed together, and logarithmic binning is applied. Straight lines represent a power law distribution fit to the degree distribution on a log–log scale. The fifth column presents the regression results from three different domains and metagenomes. The distribution and fitting results show that superlinear classes (EC1 and EC3 + EC4) have degree distributions with the longest and heaviest tails across all domains and metagenomes.

contingency. If the trade-off between optimization and contingency is a general feature of biochemical organization, it presents a counterintuitive approach to searching for the universal laws that could govern all biochemical systems; rather than focusing on universal components, it suggests that we should instead focus efforts on cases where there is maximal diversity in component membership as it is in these cases where we are most likely to observe optimization toward the hard physical limits that could apply to any biochemical system.

Our results have implications for understanding generalizable features of biochemistry that can inform generic constraints present at the origin of life on Earth and are relevant to searches for other examples of life, including life in alien environments or synthetically designed life. This is because the scaling laws arise due to the interactions of hundreds of chemical compounds interconverted in biochemical networks and seem to depend on bulk features of how groups of enzymes contribute to overall network architecture. Thus, we can conjecture that other examples of life might be subject to the same universal constraints, which arise due to physical limitations on the architecture of complex webs of chemical reactions. Steps to validate the scaling laws identified herein as truly universal should include future work focused on uncovering more about the underlying mechanisms. For example, it is important to understand if the observed exponents can in fact be directly derived from topology and whether they emerge from processes that are easily generalizable and connected with physical laws or are the result of specific and contingent evolutionary trajectories. Most of the biological scaling relationships observed previously have been connected to the former (5), suggesting that there is hope that the biochemical universality that we have observed here is likewise constrained by physics and can also be expected to be truly universal. Building mechanistic theory to explain these scaling behaviors would also allow for determining whether certain exponents are indistinguishable from one another and would help us to assess if the mechanisms are particular to life on Earth. If the identified scaling laws are indeed universal, they can also provide new frameworks for constraining inferences about the most ancient forms of life on Earth and be used to predict missing enzyme diversity in the biosphere, including within specific functional classes and domains. They can provide new constraints on revised models for LUCA, as we have proposed here, or be used to provide broad constraints on underannotation and missing functions in genomic and metagenomic data. However, we do not know how the observed exponents might have changed over evolutionary time nor whether they have converged to universal values. Scaling analyses are often restricted to extant life where enough data can be gathered to verify scaling exponents. Future studies should aim to find ways to verify the changes in these exponents over evolutionary timescales in order to answer questions about ultimate universality.

Overall, our analyses indicate that it is possible to analyze questions of biochemical universality from the perspective of statistical regularities and macroscale patterns. This opens new avenues of research into features of biochemical universality that can extend to examples not currently accessible based on traditional notions of universal biochemistry focused strictly on the exact identity of component compounds and molecules. Such advances in making statistical predictions will become increasingly important for astrobiology (31) as there currently exist no frameworks allowing quantitative predictions about the earliest biochemistries at the start of life on Earth nor the biochemistries that are possible on other worlds.

Materials and Methods

Acquiring Genomic and Metagenomic Data from the Joint Genome Institute. Using a text-mining Python script, we retrieved metadata, genome statistics data, and EC lists for samples from the DOE-JGI IMG/M database. DOE-JGI IMG/M is a comparative genomics database that contains genetic and biochemical data, including archaea, bacteria, eukarya, and metagenomes, among others (10, 32). Datasets were acquired between 18 June and 27 June 2019. Genomes and metagenomes hosted by DOE-JGI IMG/M are divided into the Joint Genome Institute (JGI) and all subcategories. These subcategories refer to where the sample was sequenced. For our dataset, archaea and eukarya are from the all category, and bacteria and metagenomes are from JGI. All was selected for archaea and eukarya so as to maximize the amount of representatives for these domains, whereas JGI was selected for bacteria and metagenomes because sufficient representation was not a concern for these groups, and we, therefore, elected to select for consistent annotation across datasets. As our analyses show, we do not see dramatically different behavior between archaea/eukaryota and bacteria/metagenomes and therefore, can conclude that the selection of the all vs. JGI is at a level of detail that does not affect the bulk trends we report here. Our dataset before filtering included 1,960 archaea samples, 16,116 bacteria samples, 677 eukarya samples, and 21,667 metagenomic samples. Archaea and bacteria samples came in the form of isolates, single-amplified genomes, and metagenome-assembled genomes. Eukarya samples came strictly from isolates. Metagenomes came primarily from environmental samples (*SI Appendix*, Fig. S1) but also include data from host-associated and engineered environments. For each sample, we pulled general study metadata (which are originally amassed in the Genome OnLine Database and adhere to metadata standards as defined by the Genomics Standards Consortium) (2), genome statistics (e.g., the total number of base pairs, the total number of genes, and the number of protein-coding genes, etc.), and a list of Enzyme Commission numbers.

Data Filtering. To make the annotation quality of the samples consistent across our dataset for our analyses, we selected a subset of the dataset available from JGI through the following steps.

In the Enzyme Commission number assignment pipeline, enzyme functions are assigned as a subset of protein-coding genes with function assignment, so the number of genes assigned an Enzyme Commission number should necessarily always be less than or equal to the number of protein function assignments. Samples that do not satisfy this condition were removed from our dataset.

Additionally, a significant portion of the metagenomic samples (~1/3) had functional assignments for 100% of their protein-coding genes. Complete functional annotations do not exist for even the best-studied organisms, such as *Escherichia coli* and yeast. We, therefore, removed metagenomic samples with 100% functional annotation from the dataset as these are likely produced in error and are overannotated.

Finally, we cleaned the dataset by removing samples from the dataset that fell under a threshold for their number of genes and their genome size (in base pairs). Thresholds were determined by searching for information about gene counts and genome size in minimal genomes/organisms. For archaea and bacteria, we removed samples with fewer than 1,364 genes, which is reflective of the sizes of the smallest free-living prokaryotes, from which we selected *Pelagibacter ubique* with 1,354 genes (33) as a representative size. For eukarya, we removed samples with fewer than 4,718 genes based on the size of one of the smallest known free-living eukaryotes, *Ashbya gossypii* (34).

Unlike genome datasets, metagenomes were a bit more complicated for the filtering since there is no coherent concept of genome length for metagenomes or of a "minimal" metagenome. We removed metagenomes with fewer than 20,000 genes, leaving space for the smallest metagenomes to hypothetically contain some 10 to 20 individual archaea or bacteria.

The final product was a dataset containing 1,194 archaea (a 36.5% reduction), 10,434 bacteria (a 29.4% reduction), 267 eukarya (a 60.5% reduction), and 6,112 metagenomes (a 44.6% reduction) (*SI Appendix*, Table S3 shows the statistics on the initial and cleaned datasets, and *SI Appendix*, Figs. S2–S5 shows statistical distributions of raw and filtered data).

Consensus LUCA Enzyme Functions. To identify consensus enzyme function predictions for the LUCA, the results of eight previously published LUCA genome studies (35–42) were mapped onto clusters within the EggNOG database (43). LUCA genome predictions from all eight studies were mapped onto UniProt accessions (44), as in the LUCApedia database (16). These UniProt accessions represent individual proteins, but the genome content of LUCA is more appropriately represented in modern taxa as larger protein families. As such, the UniProt accessions corresponding to the results of each LUCA genome study were mapped onto protein families in the EggNOG database by way of the file <uniprot-15-May-2015.LUCA.tsv> downloaded from the EggNOG site. Any protein family predicted by four or more of the eight LUCA genome studies was retained as a consensus LUCA protein family, which resulted in 366 such families.

An ancestral Enzyme Commission number was subsequently inferred for each consensus LUCA protein family. First, the Enzyme Commission numbers associated with each consensus LUCA protein family were identified through the annotations of their component proteins found in the UniProt database. Only reviewed UniProt accessions were considered in this analysis (45). Of the 366 consensus LUCA protein families, 310 contained at least one reviewed UniProt accession with an associated Enzyme Commission number. In order to infer whether a given enzyme function was ancestral to the protein family, the associated taxonomic identifications of each UniProt accession in an EggNOG family were used to determine how common the enzyme function was across the three domains of life. Taxonomic identifications were acquired from the National Center of Biotechnology Information (NCBI). If an enzyme function was only predicted in one taxonomic domain, it was not included in the final list of consensus LUCA Enzyme Commission numbers. If a single EggNOG cluster contained more than one associated Enzyme Commission number, only the Enzyme Commission numbers with the broadest taxonomic range were retained. The resulting list of consensus LUCA enzyme functions contains 200 EC numbers from 199 EggNOG clusters. *SI Appendix* contains the consensus LUCA EggNOG clusters and their predicted ancestral Enzyme Commission number(s).

Fitting Scaling Laws to Empirical Data. Power laws are the natural way to address features that have consistent relationships over large changes in scale. This is in contrast to the linear or polynomial fits often used in molecular biology to capture the interconnection of features. It should be noted that a power law fit reduces to a linear fit when the exponent is equal to one. We provide ordinary least squares fits for both the linear and power law fits to the data in *SI Appendix*. Our analyses show that the power laws are a consistently better fit to the data and typically have exponents that are distinguishable from one, thus motivating our discussion of power law fits.

Determining Universality of Enzyme Functions, Reactions, and Compounds. To quantify universality across enzyme functions, reactions, and compounds, we calculated AUC scores for the corresponding ranked-frequency distribution curve of each respective function, reaction, or compound across each dataset. We applied Simpson's rule for the calculation of AUC score. Simpson's rule, or a three-point rule Newton–Cotes formula, is one of the techniques of numerical integration using a piecewise quadratic polynomial for approximating the area under a given arbitrary curve. We used the module scipy.integrate.simps of the Python package SciPy for the implementation of Simpson's rule.

Network Expansion. Network expansion was performed to look for systematic patterns in the emergence of ECs. Network expansion is an algorithm where a small set of compounds—referred to here as seed sets—is input into a biochemical network. This network is organized so that compounds are nodes, while reactions are edges, thereby linking substrates and reactions. If the compounds in the seed set make up a full substrate list of a given reaction, then that reaction is considered possible, and all products of that reaction are added to the growing seed set. This process of performing reactions with the compounds available in the seed set is then repeated until no new compounds can be added (21). Here, we used an initial starting seed set of six compounds, which were likely readily available in the primordial ocean—H_2O, CO_2, H_2SO_4, H_3PO_4, NH_3, and H^+—for the primordial expansion (23). For the random expansion, we chose 1,000 sets of six different randomly selected biochemical compounds.

Enzyme–Enzyme Network Construction and Centrality Analysis. We constructed unipartite networks for each annotated genome from archaea,

bacteria, and eukaryota taxa and from annotated metagenomes. Each enzyme–enzyme network consists of enzymes annotated in the corresponding genome or metagenome, and two enzymes are connected to each other when their biochemical dependence can be projected onto the space of enzymes in the way that a product of a reaction catalyzed by one enzyme is used to a substrate of another reaction catalyzed by the other enzyme. We also analyzed the degree centrality and betweenness centrality of all enzyme–enzyme networks. Degree centrality of an enzyme is defined as the number of enzymes connected to the enzyme, which indicates the local functional impact of the enzyme. The degree distributions of biochemical networks are well known as heavy-tailed distributions compared with the exponential degree distribution of random networks. To measure the heaviness of the degree distribution, we utilized a power law distribution, one of the simplest heavy-tailed distributions, by performing linear regression on the degree distribution on a log–log scale. The betweenness centrality of a node is defined as the fraction of the number of shortest paths connecting every pair of nodes in a network over the number of those paths going through the given node. Hence, high betweenness centrality often identifies the most influential nodes in the network, the ones that control the large-scale connected paths and tend to bridge highly clustered parts in the network.

Data Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by funding from John Templeton Foundation Grant 61184 (to D.C.G., B.K., J.M., H.K., and S.I.W.), NSF Grant 1840301 (to C.P.K.), NASA Grant 80NSSC18K1140 (to C.P.K. and S.I.W.), and NASA Grant GR40991 (to H.K. and S.I.W.).

1. N. R. Pace, The universal nature of biochemistry. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 805–808 (2001).
2. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).
3. N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (CRC Press, 2018).
4. G. B. West, *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies* (Penguin Books, 2017).
5. C. P. Kempes, M. A. R. Koehl, G. B. West, The scales that limit: The physical boundaries of evolution. *Front. Ecol. Evol.* **7**, 242 (2019).
6. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabási, The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
7. H. Kim, H. B. Smith, C. Mathis, J. Raymond, S. I. Walker, Universal scaling across biochemical networks on Earth. *Sci. Adv.* **5**, eaau0149 (2019).
8. International Union of Biochemistry and Molecular Biology. Nomenclature Committee, *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes* (Academic Press, 1992).
9. S. Mukherjee *et al.*, Genomes OnLine Database (GOLD) v.8: Overview and updates. *Nucleic Acids Res.* **49** (D1), D723–D733 (2021).
10. I. A. Chen *et al.*, IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47** (D1), D666–D677 (2019).
11. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1460–E1469 (2017).
12. A. M. Makarieva *et al.*, Mean mass-specific metabolic rates are strikingly similar across life's major domains: Evidence for life's metabolic optimum. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 16994–16999 (2008).
13. C. P. Kempes, S. Dutkiewicz, M. J. Follows, Growth, metabolic partitioning, and the size of microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 495–500 (2012).
14. 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).
15. C. P. Kempes, L. Wang, J. P. Amend, J. Doyle, T. Hoehler, Evolutionary tradeoffs in cellular composition across diverse bacteria. *ISME J.* **10**, 2145–2157 (2016).
16. A. D. Goldman, T. M. Bernhard, E. Dolzhenko, L. F. Landweber, LUCApedia: A database for the study of ancient life. *Nucleic Acids Res.* **41**, D1079–D1082 (2013).
17. C. Woese, The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854–6859 (1998).
18. C. R. Woese, On the evolution of cells. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8742–8747 (2002).
19. N. Goldenfeld, C. Woese, Life is physics: Evolution as a collective phenomenon far from equilibrium. *Annu. Rev. Condens. Matter Phys.* **2**, 375–399 (2011).
20. L. M. A. Bettencourt *et al.*, The interpretation of urban scaling analysis in time. *J. R. Soc. Interface* **17**, 20190846 (2020).
21. T. Handorf, O. Ebenhöh, R. Heinrich, Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512 (2005).
22. J. Raymond, D. Segrè, The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767 (2006).
23. J. E. Goldford, H. Hartman, T. F. Smith, D. Segrè, Remnants of an ancient metabolism without phosphate. *Cell* **168**, 1126–1134.e9 (2017).
24. J. E. Goldford, H. Hartman, R. Marsland III, D. Segrè, Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. *Nat. Ecol. Evol.* **3**, 1715–1724 (2019).
25. H. B. Smith, A. Drew, J. F. Malloy, S. I. Walker, Seeding biochemistry on other worlds: Enceladus as a case study. *Astrobiology* **21**, 177–190 (2021).
26. M. E. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **64**, 016132 (2001).
27. L. E. Hays *et al.*, *Astrobiology Strategy* (National Aeronautics and Space Administration [NASA], Washington, DC, 2015).
28. V. B. Pinheiro *et al.*, Synthetic genetic polymers capable of heredity and evolution. *Science* **336**, 341–344 (2012).
29. D. A. Malyshev *et al.*, A semi-synthetic organism with an expanded genetic alphabet. *Nature* **509**, 385–388 (2014).
30. K. B. Muchowska *et al.*, Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol.* **1**, 1716–1721 (2017).
31. S. I. Walker *et al.*, Exoplanet biosignatures: Future directions. *Astrobiology* **18**, 779–824 (2018).
32. I. A. Chen *et al.*, The IMG/M data management and analysis system v.6.0: New tools and advanced capabilities. *Nucleic Acids Res.* **49** (D1), D751–D763 (2021).
33. S. J. Giovannoni *et al.*, Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
34. F. S. Dietrich *et al.*, The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**, 304–307 (2004).
35. Aj Harris, A. D. Goldman, Phylogenetic reconstruction shows independent evolutionary origins of mitochondrial transcription factors from an ancient family of RNA methyltransferase proteins. *J. Mol. Evol.* **86**, 277–282 (2018).
36. B. G. Mirkin, T. I. Fenner, M. Y. Galperin, E. V. Koonin, Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2 (2003).
37. L. Delaye, A. Becerra, A. Lazcano, The last common ancestor: What's in a name? *Orig. Life Evol. Biosph.* **35**, 537–554 (2005).
38. S. Yang, R. F. Doolittle, P. E. Bourne, Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 373–378 (2005).
39. J. A. G. Ranea, A. Sillero, J. M. Thornton, C. A. Orengo, Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.* **63**, 513–525 (2006).
40. M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, G. Caetano-Anollés, Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).
41. V. Srinivasan, H. J. Morowitz, The canonical network of autotrophic intermediary metabolism: Minimal metabolome of a reductive chemoautotroph. *Biol. Bull.* **216**, 126–130 (2009).
42. M. C. Weiss *et al.*, The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
43. J. Huerta-Cepas *et al.*, eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
44. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
45. E. Boutet *et al.*, UniProtKB/Swiss-Prot, the manually annotated section of the UniProt knowledgebase: How to use the entry view. *Methods Mol. Biol.* **1374**, 23–54 (2016).