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

Dylan C. Gagler¹, Bradley Karas¹, Chris Kempes³, John Malloy¹, Veronica Mierzejewski¹, Aaron D. Goldman⁴,⁵, Hyunju Kim¹,²,⁶*, Sara Imari Walker¹,²,³,⁵,⁶*
¹School of Earth and Space Exploration, Arizona State University, Tempe AZ USA
²Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe AZ USA
³Santa Fe Institute, Santa Fe, NM USA
⁴Department of Biology, Oberlin College, Oberlin OH USA
⁵Blue Marble Space Institute for Science, Seattle WA USA
⁶ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe AZ USA
*Hyunju Kim, Sara Walker

Email: sara.i.walker@asu.edu; hkim78@asu.edu

This PDF file includes:

- Supplementary text
- Figures S1 to S21
- Tables S1 to S11
- SI References

Other supplementary materials for this manuscript include the following:

- Datasets S1 to S5
Planetary Distribution of Metagenomic Data

Figure S1. Distribution of the 11,955 metagenomes in our filtered dataset, including their biome classification. Data from the Joint Genome Institute.¹
Enzymes are organized hierarchically, according to the specificity of their classification as established by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), see 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 (in this case the specified alcohol dehydrogenase reaction). It is important to recognize that enzyme commission numbers are assigned to enzymes to describe their function, and therefore an individual enzyme can be assigned more than one enzyme commission number depending on how many functions it has. Each of the 6 primary enzyme classes used in this study have different roles in biology, see Table S2. As of May 2020, there are 7736 EC numbers in circulation (7646 if you exclude EC7) as per the ExplorEnz database: 2402 oxidoreductases, 2085 transferases, 1783 hydrolases, 815 lyases, 326 isomerases, and 235 ligases. If one only includes current enzyme commission numbers, i.e., those that have not been transferred to different numbers or deleted, the numbers are as shown in Table S2, which provides the total number of enzymes in each EC class used for this study. A full list of accepted enzyme commission numbers is available on the official database of the Enzyme Nomenclature List, ExplorEnz.

Enzyme commission numbers are particularly appealing for this study because they classify enzymes by their reactions, as opposed to by sequence similarity or molecular structure, which offers a way to quantitatively map and categorize biochemical space from the perspective of chemical function. The primary drawbacks of the classification of enzyme functions into the hierarchy of enzyme commission numbers are that they do not include detailed information on reaction mechanisms, kinetic parameters, and small differences in sequence and protein structure. However, these drawbacks do not impact our use of enzyme commission numbers in this study since we are interested in statistical patterns in the functions that this classification codifies.

As described in the Methods, data was acquired from the Department of Energy Joint Genome Institute’s Integrated Microbial Genomes and Microbiomes (DOE-JGI IMG/M) database. IMG/M is a comparative genomics database which contains genetic and biochemical data, including archaea, bacteria, eukarya, and metagenomes, among others. Through IMG’s annotation pipeline, protein coding genes are assigned KEGG Orthology (KO) numbers, which are subsequently linked to enzyme commission number identifiers. Sequence data is either generated internally by the JGI, gathered from GenBank, or submitted by external researchers. Before being added to the database, unannotated sequences are put through JGI’s Microbial Genome Annotation Pipeline (MGAP V.4), which involves structural annotation (the identification of protein-coding genes, non-coding RNAs, regulatory RNA features and binding motifs, and CRISPR elements) and then functional annotation (function prediction of protein-coding genes). Along with alignment against different protein databases, proteins are associated with KEGG Orthology (KO) terms, from which JGI assigned enzyme commission numbers are derived.

Importantly for the current work, enzyme function can be assigned to genetic sequences via enzyme commission numbers. Once a genome or metagenome has been sequenced and protein coding regions have been identified, these protein coding regions can be referenced against protein
databases to assign their function(s). Functional annotation in this study was conducted using data in the Kyoto Encyclopedia of Genes and Genomes (KEGG)\textsuperscript{5}, where proteins are assigned KEGG Orthology (KO) that contain an internal linkage in the database to enzyme commission numbers.

We next briefly describe functions in each EC class.
Table S1. Hierarchical scheme for organizing enzyme commission numbers, as assigned by the Enzyme Commission Classification Scheme.

| Hierarchical Classification Level | EC digit | Example                                           |
|----------------------------------|----------|---------------------------------------------------|
| Class                            | 1st digit| 1.x.x.x Oxidoreductases                           |
| Sub-Class                        | 2nd digit| 1.1.x.x CH-OH groups as donors                    |
| Sub-subclass                     | 3rd digit| 1.1.1.x NAD+ or NADP+ as electron acceptors       |
| Serial Number                    | 4th digit| 1.1.1.1 alcohol dehydrogenase                     |
Table S2. Enzyme class description, and total functional diversity in each class (totals from ExploreEnz database\textsuperscript{2}) for biochemical reaction classes as defined by the primary digit of the Enzyme Commission (EC) numbers (EC 7 translocases, which transport substrates across membranes are not included as they were not included in this study).

| Class | Class Name | Function | Current Total |
|-------|------------|----------|---------------|
| 1     | Oxidoreductase | Oxidation-reduction reactions, involving transfer of electrons | 1905          |
| 2     | Transferase | Transfer of functional groups between molecules | 1917          |
| 3     | Hydrolase | Cleavage of molecular bonds via hydrolysis | 1315          |
| 4     | Lyase | Cleavage of molecular bonds via reaction mechanisms other than hydrolysis | 705           |
| 5     | Isomerase | Intramolecular rearrangement | 304           |
| 6     | Ligase | Joining of large molecules | 220           |

**Oxidoreductases.** Oxidoreductases (EC1) are an abundant and diverse class of enzymes which catalyze oxidation-reduction reactions. Oxidoreductases are critically involved in both aerobic and anaerobic respiration as membrane electron transfer proteins, helping to drive the generation of proton gradients and subsequent ATP synthesis\textsuperscript{6}. Furthermore, oxidoreductases are integrally involved in metabolite processing (via the synthesis of alcohols, ammonia, carboxylic acids, alkenes and alkanes, and breakdown of amino acids) and the uptake and elemental fixation of various inorganic compounds including sulfur and nitrogen\textsuperscript{7}.

**Transferases.** Transferases (EC2) are also abundant and diverse in the biosphere, catalyzing the transfer of functional groups between molecules. These common enzymes drive much of biology’s core anabolic work, such as peptide bond formation, nucleotide synthesis, carbohydrate synthesis, and DNA and RNA synthesis, in addition to also participating in glycolysis, the pentose phosphate pathway, and many regulatory functions such as detoxification, molecular trafficking, protein degradation, cellular energy maintenance, and cofactor coordination\textsuperscript{8}.
**Hydrolases.** Hydrolases (EC3) catalyze the hydrolytic cleavage of various types of bonds and power much of biology's catabolic work: cleaving ester bonds (which store fatty acids as glycerides), glycosidic bonds (which compose complex sugars and are structurally involved in DNA), ether bonds (which impart biomolecules with a high degree of resistance to biological mineralization), peptide bonds, and phosphate bonds, among others.

**Lyases.** Lyases (EC4) catalyze bond cleavage by means other than hydrolysis and differ stoichiometrically in how two or more substrates are used for one reaction direction. While hydrolases use water to cleave ester-type linkages, lyases tend to break or create double bonds, primarily carbon-carbon double bonds. Among other things, these enzymes are responsible for liberating, and thus rendering bioavailable, certain compounds like water, ammonia, chloride, and sulfide.

**Isomerases.** Isomerases (EC5) are the second least abundant enzyme class and catalyze intramolecular rearrangements. Isomerases are primarily involved in carbohydrate metabolism and in terpenoid/polyketide metabolism, which is important for generating secondary metabolites. Furthermore, racemases/epimerases, the primary isomerase subclass (EC 5.1.x.x), are involved in the stereochemical interconversion of amino acids, which is important for bacterial cell wall assembly and other bacterial structures and functions. Furthermore, broad spectrum racemases, which produce and release several D-amino acids to the environment, suggest the importance of isomerases in ecosystem level function.

**Ligases.** Ligases (EC 6) are anabolic enzymes that catalyze the joining of molecules, hydrolyzing a nucleotide triphosphate (NTP) such as ATP in the process. Ligases are involved in many biologically essential tasks, including DNA and RNA repair and regulation, coenzyme A manipulation, and protein synthesis (via aminoacyl-tRNA synthetases adding amino acids to tRNA molecules).
Filtered Data

Table S3. Statistics for raw versus filtered data sampled from JGI.

|        | Domain      | Initial | No enzyme | Filtered |
|--------|-------------|---------|-----------|----------|
| Raw    | Archaea     | 1960    | 0         | 1960     |
| Filtered | Archaea | 1960    | 0         | 1282     |
| Raw    | Bacteria    | 16755   | 639       | 16116    |
| Filtered | Bacteria | 16755   | 150       | 11759    |
| Raw    | Eukaryota   | 710     | 33        | 677      |
| Filtered | Eukaryota | 710     | 0         | 200      |
| Raw    | Metagenome  | 21667   | 0         | 21667    |
| Filtered | Metagenome | 21667   | 0         | 11955    |
Figure S2. Extent of enzyme annotation in raw (left) and filtered (right) data sets. Histogram plots with bins of 1% width showing the extent of enzyme annotation in our dataset prior to any data cleaning or processing on the left and after filtering on the right. The x-axis depicts the percentage of protein coding genes with enzymatic assignments out of total protein coding genes in each genome or metagenome. Enzyme assignments are made via KEGG Orthology (KO).
Figure S3. Extent of functional annotation in raw (left) and filtered (right) datasets. Histogram plots with bins of 1% width showing the extent of functional annotation in our dataset prior to any data cleaning or processing on the left and after filtering on the right. The x-axis depicts the percentage of protein coding genes with functional assignments out of total protein coding genes in each genome or metagenome. Note the substantial fraction of metagenomes with ~100% functional annotation (N = 7591 with an additional N = 11 of > 120%), which are not viable for use due to artifacts of over-annotation.
Figure S4. Extent of protein coding genes with enzymatic versus functional annotation in raw (left) and filtered (right) datasets. Scatter plot showing the extent of annotation in the dataset prior to any data processing or filtering on the left and after filtering on the right. The x-axis depicts the percentage of protein coding genes with functional annotation in each genome or metagenome and the y-axis depicts the percentage of protein coding genes with enzymatic annotation in each genome or metagenome. Note the eukarya data with enzyme annotation percentages between 0 and 25% and functional annotation percentages less than 20%. Fungi were removed from the dataset (see description in text).
Figure S5. Extent of protein coding genes with enzymatic versus functional annotation in raw (left) and filtered (right) datasets. Scatter plot showing the extent of annotation in the dataset prior to any data processing or cleaning on the left and after filtering on the right. The x-axis depicts the percentage of protein coding genes with functional annotation in each genome or metagenome and the y-axis depicts the total number of genes in the genome. Enzyme annotations are assigned via functional annotations, so y-axis values that are greater than their corresponding x-axis values are not sensible. Fungi were removed from the dataset also.
Scaling Law Fits to Empirical Data: Enzyme Commission Classification for Enzyme Classes

**Table S4.** Table of regression values for linear fits to scaling behavior of ECs across domains and metagenomes.

| Dataset       | Dataset Type   | Slope  | Bacteria | Eukaryota | Metagenome | Pan-taxa | All  |
|---------------|----------------|--------|----------|-----------|------------|----------|------|
| EC 1          | Slope          | 0.227  | 0.266    | 0.277     | 0.345      | 0.263    | 0.312|
| Oxidoreductases| Slope CI       | -0.005 | -0.001   | -0.007    | -0.001     | -0.001   | -0.001|
|               | Intercept      | -12.373| -31.474  | -31.754   | -110.285   | -28.980  | -58.987|
|               | Intercept CI   | -1.712 | -0.764   | -4.823    | -1.587     | -0.666   | -0.559|
|               | R-squared      | 0.876  | 0.929    | 0.966     | 0.970      | 0.933    | 0.983|
|               | Slope Std Error| 0.002  | 0.001    | 0.004     | 0.001      | 0.001    | 0.000|
|               | Intercept Std Error| 0.873 | 0.390 | 2.446 | 0.810 | 0.340 | 0.285 |
| EC 2          | Slope          | 0.311  | 0.292    | 0.338     | 0.294      | 0.301    | 0.299|
| Transferases  | Slope CI       | -0.006 | -0.001   | -0.010    | -0.001     | -0.001   | -0.000|
|               | Intercept      | 13.256 | 31.949   | 27.889    | 36.558     | 25.992   | 28.123|
|               | Intercept CI   | -2.057 | -0.768   | -6.616    | -1.512     | -0.717   | -0.467|
|               | R-squared      | 0.901  | 0.939    | 0.957     | 0.963      | 0.941    | 0.987|
|               | Slope Std Error| 0.003  | 0.001    | 0.005     | 0.001      | 0.001    | 0.000|
|               | Intercept Std Error| 1.049 | 0.392 | 3.355 | 0.771 | 0.366 | 0.238 |
| EC 3          | Slope          | 0.142  | 0.171    | 0.201     | 0.156      | 0.175    | 0.168|
| Hydrolases    | Slope CI       | -0.004 | -0.001   | -0.007    | 0.000      | -0.001   | 0.000|
|               | Intercept      | -10.704| -14.566  | -23.074   | 6.497      | -17.245  | -11.903|
|               | Intercept CI   | -1.566 | -0.465   | -4.518    | -0.572     | -0.433   | -0.228|
|               | R-squared      | 0.767  | 0.935    | 0.944     | 0.981      | 0.936    | 0.990|
|               | Slope Std Error| 0.002  | 0.000    | 0.003     | 0.000      | 0.000    | 0.000|
|               | Intercept Std Error| 0.798 | 0.237 | 2.291 | 0.292 | 0.221 | 0.116 |
| EC 4          | Slope          | 0.163  | 0.136    | 0.096     | 0.118      | 0.131    | 0.123|
| Lyases        | Slope CI       | -0.003 | -0.001   | -0.005    | 0.000      | -0.001   | 0.000|
|               | Intercept      | -10.991| -8.147   | -1.131    | 7.611      | -5.040   | -0.183|
|                | Intercept CI | R-squared | Slope Std Error | Intercept Std Error |
|----------------|--------------|-----------|-----------------|--------------------|
| **EC 5**       | -1.018       | 0.912     | 0.001           | 0.519              |
| **Isomerases** |              | 0.949     | 0.000           | 0.167              |
|                | -3.413       | 0.871     | 0.003           | 1.731              |
|                | -0.596       | 0.965     | 0.000           | 0.304              |
|                | -0.339       | 0.931     | 0.000           | 0.173              |
|                | -0.198       | 0.986     | 0.000           | 0.101              |
| **EC 6**       |              | 0.002     | 0.001           | 0.001              |
| **Ligases**    |              | 0.000     | 0.002           | 0.000              |
|                |              | -0.004    | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | -2.507    | 15.035          | 1.647              |
|                |              | -0.361    | 1.647           | 6.407              |
|                |              | -0.248    | 0.854           | 0.968              |
|                |              | -0.135    | 0.968           | 0.968              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
|                |              | 0.000     | 0.000           | 0.000              |
Table S5. Table of regression values for power law (log base 10) fits to scaling behavior of ECs across domains and metagenomes.

| EC 1  | Dataset       | Slope | Bacteria | Eukaryota | Metagenome | Pan-taxa | All     |
|-------|---------------|-------|----------|-----------|------------|----------|---------|
| Oxidoreductases | Slope CI | -0.023 | -0.006  | -0.037 | -0.004 | -0.006 | -0.002 |
| Oxidoreductases | Intercept | -1.169 | -1.347  | -1.570 | -1.499 | -1.317 | -1.389 |
| Oxidoreductases | Intercept CI | -0.059 | -0.017  | -0.102 | -0.014 | -0.016 | -0.007 |
| R-squared | 0.885 | 0.925  | 0.962  | 0.965  | 0.929  | 0.979  |
| Transferases | Slope CI | -0.013 | -0.003  | -0.021 | -0.003 | -0.003 | -0.001 |
| Transferases | Intercept | -0.298 | -0.091  | -0.036 | -0.214 | -0.153 | -0.183 |
| Transferases | Intercept CI | -0.034 | -0.009  | -0.059 | -0.009 | -0.009 | -0.004 |
| R-squared | 0.936 | 0.956  | 0.970  | 0.969  | 0.956  | 0.986  |
| Hydrolases | Slope CI | -0.033 | -0.006  | -0.046 | -0.003 | -0.006 | -0.002 |
| Hydrolases | Intercept | -1.458 | -1.387  | -1.759 | -0.839 | -1.514 | -1.289 |
| Hydrolases | Intercept CI | -0.084 | -0.015  | -0.127 | -0.008 | -0.016 | -0.006 |
| R-squared | 0.795 | 0.937  | 0.944  | 0.979  | 0.929  | 0.977  |
| Lyases | Slope CI | -0.022 | -0.005  | -0.046 | -0.003 | -0.005 | -0.002 |
| Lyases | Intercept | -1.657 | -1.355  | -1.068 | -0.892 | -1.258 | -1.053 |
|                  | Intercept | CI         | Slope | Std Error | R-squared | Slope | Std Error | Intercept | Std Error |
|------------------|-----------|------------|-------|-----------|-----------|-------|-----------|-----------|-----------|
|                  | -0.057    | -0.014     | -0.127| -0.010    | -0.014    | -0.006|           |           |           |
| EC 5             | 0.911     | 0.944      | 0.906 | 0.972     | 0.929     | 0.977 |           |           |           |
| Isomerases       | 0.011     | 0.003      | 0.023 | 0.002     | 0.003     | 0.001 |           |           |           |
|                  | 0.029     | 0.007      | 0.065 | 0.005     | 0.007     | 0.003 |           |           |           |
| Ligases          | 0.820     | 0.958      | 0.959 | 0.887     | 0.944     | 0.921 |           |           |           |
|                  | -0.028    | -0.006     | -0.066| -0.004    | -0.006    | -0.002|           |           |           |
|                  | -0.718    | -1.064     | -1.263| -0.861    | -1.029    | -0.967|           |           |           |
|                  | -0.072    | -0.017     | -0.185| -0.013    | -0.017    | -0.007|           |           |           |
|                  | 0.714     | 0.885      | 0.803 | 0.938     | 0.867     | 0.959 |           |           |           |
|                  | 0.015     | 0.003      | 0.034 | 0.002     | 0.003     | 0.001 |           |           |           |
|                  | 0.037     | 0.009      | 0.094 | 0.007     | 0.009     | 0.004 |           |           |           |
| EC 6             | 0.733     | 0.722      | 0.462 | 0.573     | 0.675     | 0.559 |           |           |           |
|                  | -0.021    | -0.006     | -0.032| -0.003    | -0.006    | -0.002|           |           |           |
|                  | -0.147    | -0.197     | 0.466 | 0.189     | -0.063    | 0.242 |           |           |           |
|                  | -0.052    | -0.016     | -0.088| -0.010    | -0.016    | -0.006|           |           |           |
|                  | 0.794     | 0.834      | 0.807 | 0.914     | 0.801     | 0.907 |           |           |           |
|                  | 0.010     | 0.003      | 0.016 | 0.002     | 0.003     | 0.001 |           |           |           |
|                  | 0.027     | 0.008      | 0.045 | 0.005     | 0.008     | 0.003 |           |           |           |
**Table S6.** Table of regression values for power law (log base e) fits to scaling behavior of ECs across domains and metagenomes.

| Dataset | Slope | Archaea | Bacteria | Eukaryota | Metagenome | Pan-taxa | All |
|---------|-------|---------|----------|-----------|------------|----------|-----|
| EC 1    | Slope | 1.175   | 1.239    | 1.327     | 1.291      | 1.229    | 1.255 |
| Oxidoreductases | Slope CI | -0.023   | -0.006   | -0.037    | -0.004     | -0.006   | -0.002 |
|          | Intercept | -2.692   | -3.101   | -3.616    | -3.452     | -3.032   | -3.197 |
|          | Intercept CI | -0.136   | -0.040   | -0.236    | -0.032     | -0.036   | -0.015 |
|          | R-squared | 0.885    | 0.925    | 0.962     | 0.965      | 0.929    | 0.979 |
|          | Slope Std Error | 0.012    | 0.003    | 0.019     | 0.002      | 0.003    | 0.001 |
|          | Intercept Std Error | 0.069    | 0.020    | 0.120     | 0.016      | 0.018    | 0.008 |
| EC 2    | Slope | 0.937   | 0.868    | 0.864     | 0.911      | 0.890    | 0.901 |
| Transferases | Slope CI | -0.013   | -0.003   | -0.021    | -0.003     | -0.003   | -0.001 |
|          | Intercept | -0.687   | -0.209   | -0.082    | -0.492     | -0.353   | -0.421 |
|          | Intercept CI | -0.079   | -0.021   | -0.135    | -0.021     | -0.020   | -0.009 |
|          | R-squared | 0.936    | 0.956    | 0.970     | 0.969      | 0.956    | 0.986 |
|          | Slope Std Error | 0.007    | 0.002    | 0.011     | 0.001      | 0.002    | 0.001 |
|          | Intercept Std Error | 0.040    | 0.011    | 0.069     | 0.011      | 0.010    | 0.005 |
| EC 3    | Slope | 1.195   | 1.196    | 1.344     | 1.015      | 1.241    | 1.158 |
| Hydrolases | Slope CI | -0.033   | -0.006   | -0.046    | -0.003     | -0.006   | -0.002 |
|          | Intercept | -3.357   | -3.194   | -4.051    | -1.931     | -3.487   | -2.968 |
|          | Intercept CI | -0.194   | -0.035   | -0.293    | -0.019     | -0.036   | -0.015 |
|          | R-squared | 0.795    | 0.937    | 0.944     | 0.979      | 0.929    | 0.977 |
|          | Slope Std Error | 0.017    | 0.003    | 0.023     | 0.001      | 0.003    | 0.001 |
|          | Intercept Std Error | 0.099    | 0.018    | 0.149     | 0.010      | 0.018    | 0.008 |
| EC 4    | Slope | 1.303   | 1.158    | 1.014     | 0.995      | 1.124    | 1.047 |
| Lyases  | Slope CI | -0.022   | -0.005   | -0.046    | -0.003     | -0.005   | -0.002 |
|          | Intercept | -3.816   | -3.119   | -2.459    | -2.055     | -2.896   | -2.425 |
|          | Intercept CI | -0.130   | -0.032   | -0.293    | -0.022     | -0.033   | -0.013 |
|                | R-squared | 0.911 | 0.944 | 0.906 | 0.972 | 0.929 | 0.977 |
|----------------|-----------|-------|-------|-------|-------|-------|-------|
| Slope Std Error | 0.011     | 0.003 | 0.023 | 0.002 | 0.003 | 0.001 |
| Intercept Std Error | 0.066 | 0.016 | 0.149 | 0.011 | 0.017 | 0.007 |
| EC 5             | Slope     | 0.820 | 0.958 | 0.959 | 0.887 | 0.944 | 0.921 |
| Isomerases       | Slope CI  | -0.028 | -0.006 | -0.066 | -0.004 | -0.006 | -0.002 |
|                 | Intercept | -1.653 | -2.450 | -2.909 | -1.982 | -2.369 | -2.228 |
|                 | Intercept CI | -0.166 | -0.039 | -0.426 | -0.030 | -0.039 | -0.016 |
|                 | R-squared  | 0.714 | 0.885 | 0.803 | 0.938 | 0.867 | 0.959 |
|                 | Slope Std Error | 0.015 | 0.003 | 0.034 | 0.002 | 0.003 | 0.001 |
|                 | Intercept Std Error | 0.085 | 0.020 | 0.216 | 0.015 | 0.020 | 0.008 |
| EC 6             | Slope     | 0.733 | 0.722 | 0.462 | 0.573 | 0.675 | 0.559 |
| Ligases          | Slope CI  | -0.021 | -0.006 | -0.032 | -0.003 | -0.006 | -0.002 |
|                 | Intercept | -0.340 | -0.454 | 1.073 | 0.435 | -0.144 | 0.558 |
|                 | Intercept CI | -0.120 | -0.036 | -0.202 | -0.023 | -0.036 | -0.015 |
|                 | R-squared  | 0.794 | 0.834 | 0.807 | 0.914 | 0.801 | 0.907 |
|                 | Slope Std Error | 0.010 | 0.003 | 0.016 | 0.002 | 0.003 | 0.001 |
|                 | Intercept Std Error | 0.061 | 0.019 | 0.103 | 0.012 | 0.018 | 0.008 |
Scaling Law Fits to Empirical Data: Combining Lyases and Hydrolases

Lyases did not demonstrate universality across all biological groups. To investigate why, we hypothesized that lyases may not be the best functional classification. Given that both lyases and hydrolases break down molecules—albeit through different mechanisms—we analyzed these ECs together to examine their combined scaling behavior. The combined fits demonstrate consistent with super linear behavior across domains, although metagenomes show a slope close to linear.

**Table S7:** Table of regression values for power law (log base 10) fits to scaling behavior for combined hydrolases (EC 3) and lyases (EC 4).

| Dataset        | Archaea | Bacteria | Eukaryota | Metagenome |
|----------------|---------|----------|-----------|------------|
| ECs 3 and 4    |         |          |           |            |
| Slope          | 1.232   | 1.168    | 1.206     | 1.004      |
| Hydrolases and | Slope CI | 0.018    | 0.004     | 0.028      | 0.002      |
| Lyases         | Intercept | -1.210   | -1.042    | -1.170     | -0.557     |
|                | Intercept CI | 0.046    | 0.010     | 0.078      | 0.006      |
|                | R-squared | 0.934    | 0.971     | 0.973      | 0.99       |
|                | Slope Std Error | 0.009    | 0.002     | 0.014      | 0.001      |
|                | Intercept Std Error | 0.023    | 0.005     | 0.039      | 0.003      |
**Additional Scaling Plots including Pan-taxa, All Data Scaling and Combined Functional Classes**

In the manuscript we report scaling laws for the individual domains and metagenomes. Also of interest is the scaling behavior of all individuals (combining data from all three domains) or the scaling behavior of all biochemical systems in our ensemble (combining data from all three domains and metagenomes). These scaling trends are shown in Figure S6.

![Graph showing enzyme function scaling](image)

**Figure S6.** Scaling behaviors in enzyme functions for each EC, including scaling for data from all three domains taken together (Pan-taxa) and all domain data together with metagenomes (All). Biosphere values are 5440 total number of functions with an enzyme class breakdown of 1805 oxidoreductases, 1733 transferases, 761 hydrolases, 661 lyases, 271 isomerases, and 209 ligases.
We also include here scaling plots with histograms showing the density of each dataset along the x and y-axes for each EC class, as shown in Figures S7 - S12.

**Figure S7.** Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for oxidoreductases. Marginal histogram plots show counts with bin width of 1.
**Figure S8.** Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for transferases. Marginal histogram plots show counts with bin width of 1.
Figure S9. Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for hydrolases. Marginal histogram plots show counts with bin width of 1.
Figure S10. Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for lyases. Marginal histogram plots show counts with bin width of 1.
Figure S11. Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for isomerases. Marginal histogram plots show counts with bin width of 1.
Figure S12. Distribution of Bacteria, Archaea, and Eukaryota taxa and metagenome data used for scaling plots in Figure 2 of paper for ligases. Marginal histogram plots show counts with bin width of 1.
**Combined Functional Classes.** Along with examining each EC individually among biological groups, we examined both enzyme functions EC 3 and 4, as both represent functions that break down molecules and thus, may exhibit unique scaling behavior.

**Figure S13.** Combined scaling behavior in enzyme functions for EC 3 and 4, demonstrating possible universality among functions that break down molecules. Left: linear plot. Right: log (base 10) plot with ordinary least squares fit. See Table S7 for fits.
Scaling Laws and Divergence times

**Table S9.** Comparison of estimated time of evolutionary emergence for each domain from the literature, compared to the logarithmic scaling law fits (95% confidence intervals) reported in Table 1. Shading of cells indicates trends in scaling law fits, such that the smallest value is white and the highest dark grey. If two fits are not statistically distinguishable, they are colored the same color. There are two general trends: ECs whose values roughly increase with evolutionary divergence time and those whose values roughly decrease. The oxidoreductases, hydrolases and isomerases are in the first group, while transferases, lyases and ligases are in the latter. While these groupings do not correlate with scaling behaviors (superlinear, linear and sublinear), they do confirm at a very coarse level that evolutionary divergence times could be driving consistent increase or decrease in scaling coefficients depending on function.

| Domain   | Est. Time of Divergence | Oxidoreductases | Transferases | Hydrolases | Lyases | Isomerases | Ligases |
|----------|--------------------------|-----------------|--------------|------------|--------|------------|--------|
| Archaea  | 4.1 – 1.6 Ga^{12,13,14,15,16,17} | [1.152, 1.199] | [0.924, 0.951] | [1.162, 1.228] | [1.281, 1.325] | [0.792, 0.849] | [0.712, 0.753] |
| Bacteria | 3.5 – 3.2 Ga^{18,19} | [1.233, 1.245] | [0.864, 0.871] | [1.191, 1.202] | [1.153, 1.163] | [0.952, 0.964] | [0.716, 0.728] |
| Eukaryota| 2.5 – 0.8 Ga^{20,21,22,23,24} | [1.290, 1.364] | [0.843, 0.886] | [1.299, 1.390] | [1.068, 1.060] | [0.892, 1.025] | [0.430, 0.493] |
Universality of Enzyme Functions, Reactions and Compounds

The total number of enzyme functions, reactions and compounds in the filtered data used to calculate AUC scores are in Table S8.

Table S8. Table of total number of unique ECs, reactions and compounds for each dataset after filtering.

|                | Enzyme functions | Reactions | Compounds |
|----------------|------------------|-----------|-----------|
| Archaea        | 1610             | 3027      | 2826      |
| Bacteria       | 2426             | 4259      | 3761      |
| Eukaryota      | 2103             | 3971      | 3628      |
| Metagenome     | 2995             | 5167      | 4536      |
| Pan-taxon       | 2911             | 5040      | 4425      |
| ALL             | 2998             | 5171      | 4540      |
Figure S14. Top panel: Universality curves for enzymes, reactions and compounds in modern life show that compounds tend to be more universal than reactions, and reactions than enzyme functions, and that each is more universal in metagenomes than in samples from the three domains. Comparing the distribution of components of LUCA across modern life demonstrates that most LUCA compounds, reactions and functions are found in all metagenomes but are less universally distributed in individuals.
Table S10. Area Under the Curve (AUC) scores for data shown in Figure 4 top panel. A score of AUC=1.0 means all components in that category occur in 100% of samples and a value of 0.0 indicates components are found in none of the samples. Thus, AUC scores closer to 1 indicate more universality in the distribution of specific functions in a given category.

| Dataset      | Enzyme | Reaction | Compound |
|--------------|--------|----------|----------|
| Archaea      | 0.220  | 0.232    | 0.307    |
| Bacteria     | 0.226  | 0.266    | 0.335    |
| Eukaryota    | 0.300  | 0.363    | 0.421    |
| Metagenome   | 0.472  | 0.521    | 0.567    |
Correlation between Scaling Data and Universality

To determine the correlation between universality of component membership (enzyme functions, reactions, compounds) and how tightly constrained the identified EC scaling laws are, we plotted variation in the fitted value of the scaling coefficients across taxa, against the average AUC score of the three domains. Variation was calculated by taking the difference between the largest fitted value for the scaling coefficient and the smallest. The results are shown in Figure S14, which demonstrate that the distribution of enzyme functions across the data is not correlated with how tightly constrained scaling coefficients are across data sets.

Figure S15. Variation in the scaling coefficient across taxa datasets (measuring the difference between the largest and smallest coefficient value across the domains) does not correlate with the average Area Under the Curve universality score when compared across the enzyme classes. Bars in the plot above show the 95% confidence interval for the scaling coefficient for each domain and pan-taxis. The lack of correlation suggests that variation in scaling behavior is not directly correlated with universality of individual enzyme commission number identifiers within an Enzyme Class (EC).
Figure S16. Enzyme functions in a consensus model for the Last Universal Common Ancestor (LUCA), including projections based on the combination of hydrolases and lyases.
Statistics on EC Diversity Across Datasets

The percentage of each EC in each data set is shown in Figure S17 (averages and standard deviation for datasets that are ensembles of genomes or metagenomes).

Figure S17. Percentage of ECs for the LUCA model and the Biosphere, as well as averages over ensembles for each of the data sets sampled from the three domains and metagenomes. Pan-taxa data is the ensemble average over data from the three domains. From left to right the total number of unique enzyme functions for each dataset is: 154 in the LUCA model, 1610 in archaea, 2426 in bacteria, 2103 in eukaryota, 2911 in pan-taxa, 2995 in metagenomes, and 5440 in biosphere. The data show a general trend that certain EC classes become enriched, and others depleted as the total diversity of enzymes in a biochemical system increases (e.g., confirms that the scaling trends we see are held up in terms of averages for datasets).
Additional Network Expansion Results

**Primordial Expansions**

**Figure S18.** Enzyme class counts over network expansion using compounds readily available in the primordial ocean (H₂O, CO₂, H₂SO₄, H₃PO₄, NH₃, and H⁺). Hydrolases and Lyases are counted separately.
Figure S19. Enzyme class counts from 1000 network expansions with seed sets randomly chosen from the pool of all biochemical compounds. Each seed set consisted of six compounds, to compare with network expansion using six primordially available compounds. Opaque lines are the totals of individual expansions. Bold lines are the average new and cumulative compounds at each generation, respectively. The top panel includes the six Enzyme Commission Class Identifiers, whereas the bottom panel combines Hydrolases and Lyases. As with the primordial expansion shown in the main text, random expansions clearly show a trend where contributions to the network expansion process taper off roughly in order of scaling coefficient for that class when Hydrolases and Lyases are combined.
Additional Enzyme Network Results

![Enzyme Network Results](image-url)
Figure S20. Probability distribution of degree centrality (Top) and node betweenness centrality (Bottom) for functional classes of enzymes within enzyme-enzyme networks across biochemical datasets. Each column is associated with datasets from annotated genomes sampled from Archaea, Bacteria and Eukaryota taxa, Pantaxa data (three domains that are combined together), and from annotated metagenomes, respectively. For both plots, Hydrolases and Lyases are computed together, and logarithmic binning is applied.
Figure S21. Probability distribution of degree centrality (Top) and node betweenness centrality (Bottom) for functional classes of enzymes within enzyme-enzyme networks across biochemical datasets. Each column is associated with datasets from annotated genomes sampled from Archaea, Bacteria and Eukaryota taxa, Pantaxa data (three domains that are combined together), and from annotated metagenomes, respectively. For both plots logarithmic binning is applied.
Table S11. Regression values fit for the degree distribution shown in Figure 7 as 95% confidence intervals. The degree distribution was fit to the power law relation $y = ax^k$, where $k$ is the slope and $a$ is the intercept of the logarithmically-transformed regression.

| Enzyme Class   | Archaea       | Bacteria       | Eukaryota      | Metagenome    |
|----------------|---------------|----------------|----------------|---------------|
| Oxidoreductases| [-1.354, -1.091] | [-1.205, -0.793] | [-0.661, -0.352] | [-0.387, -0.094] |
| Transferases   | [-1.300, -0.973] | [-1.578, -1.159] | [-1.170, -0.877] | [-1.148, -0.933] |
| Hydrolases + Lyases | [-1.093, -0.615] | [-0.999, -0.466] | [-0.920, -0.427] | [-0.720, -0.267] |
| Isomerases     | [-3.677, 0.245] | [-2.169, -0.115] | [-3.117, -1.074] | [-2.753, -1.901] |
| Ligases        | [-1.791, -1.331] | [-1.476, -1.074] | [-1.345, -0.933] | [-0.868, -0.488] |
SI References

1. Webb, E. C. & Others. *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).

2. McDonald, A. G., Boyce, S. & Tipton, K. F. ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* 37, D593–7 (2009).

3. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* 47, D666–D677 (2019).

4. Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand. Genomic Sci.* 10, 86 (2015).

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

6. Hägerhäll, C. Succinate: quinone oxidoreductases. Variations on a conserved theme. *Biochim. Biophys. Acta* 1320, 107–141 (1997).

7. Sellés Vidal, L., Kelly, C. L., Mordaka, P. M. & Heap, J. T. Review of NAD(P)H-dependent oxidoreductases: Properties, engineering and application. *Biochim. Biophys. Acta: Proteins Proteomics* 1866, 327–347 (2018).

8. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* 47, D542–D549 (2019).

9. White, G. F., Russell, N. J. & Tidswell, E. C. Bacterial scission of ether bonds. *Microbiol. Rev.* 60, 216–232 (1996).

10. Lipmann, F. Metabolic generation and utilization of phosphate bond energy. *Adv. Enzymol. Relat. Areas Mol. Biol.* 1, 99–162 (1941).

11. Cuesta, S. M., Rahman, S. A. & Thornton, J. M. Exploring the chemistry and evolution of the isomerases. *Proc. Natl. Acad. Sci. U. S. A.* 113, 1796–1801 (2016).

12. Martinez Cuesta, S., Furnham, N., Rahman, S. A., Sillitoe, I. & Thornton, J. M. The evolution of enzyme function in the isomerases. *Curr. Opin. Struct. Biol.* 26, 121–130
13. Hernández, S. B. & Cava, F. Environmental roles of microbial amino acid racemases. *Environ. Microbiol.* 18, 1673–1685 (2016).

14. Holliday, G. L., Rahman, S. A., Furnham, N. & Thornton, J. M. Exploring the biological and chemical complexity of the ligases. *J. Mol. Biol.* 426, 2098–2111 (2014).