Received: 17 January 2021 | Revised: 5 February 2021 | Accepted: 9 February 2021

DOI: 10.1002/pro.4041

FOR THE RECORD

THE PROTEIN SOCIETY WILEY

Quantifying the taxonomic bias in enzymology

Chelsea J. Vickers[1] | **Dean Fraga**[2] | **Wayne M. Patrick**[1]

[1]Centre for Biodiscovery, School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand

[2]Department of Biology, The College of Wooster, Wooster, Ohio

Correspondence
Wayne M. Patrick, School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand.
Email: wayne.patrick@vuw.ac.nz

The copyright line for this article was changed on 19 April, 2021 after original online publication.

Abstract

The ongoing biotechnological revolution is rooted in our knowledge of enzymes. However, metagenomics is showing how little we know about Earth's enzyme repertoire. Deep sequencing has revolutionized our view of the tree of life. The genomes of newly-discovered organisms are replete with novel sequences, emphasizing the trove of enzyme structures and functions waiting to be explored by biochemists. Here, we sought to draw attention to the vastness of the "enzymatic dark matter" within the tree of life by placing enzymological knowledge in the context of phylogeny. We used kinetic parameters from the BRaunschweig ENzyme DAtabase (BRENDA) as our proxy for enzymological knowledge. Mapping 12,677 BRENDA entries onto the phylogenetic tree revealed that 55% of these data were from eukaryotes, even though they are the least diverse part of the tree. At the next taxonomic level, only four of 18 archaeal phyla and 24 of 111 bacterial phyla are represented in the BRENDA dataset. One phylum, the Proteobacteria, accounts for over half of all bacterial entries. Similarly, the supergroup Amorphea, which includes animals and fungi, contains over half the data on eukaryotes. Many major taxonomic groups are notable for their complete absence from BRENDA, including the ultra-diverse bacterial Candidate Phyla Radiation. At the species level, five mammals (including human) contribute 15% of BRENDA entries. The taxonomic bias in enzymology is strong, but in the era of gene synthesis we now have the tools to address it. Doing so promises to enrich our biochemical understanding of life and uncover powerful new biocatalysts.

KEYWORDS

archaeal and bacterial phyla, enzymatic dark matter, eukaryotic supergroups, kinetic parameters, phylogenetics, tree of life

1 | INTRODUCTION

It has now been over 120 years since Eduard Buchner famously founded the field of enzymology by reporting that a cell-free extract of yeast could "institute fermentation of carbohydrates".[1,2] In the ensuing decades, enzymologists have learned an enormous amount about the structures, functions and mechanisms of thousands of enzymes.[3] Increasingly, this fundamental knowledge is allowing us to engineer or evolve enzymes into powerful biocatalysts for industry and synthetic biology.[3,4]

Nevertheless, it is becoming apparent how little we actually know about the diversity of enzymes on Earth. Even the most biochemically well-characterized organisms – *Homo sapiens*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* – have had turnover numbers (k_{cat} values) reported for fewer than 10% of their enzymes.[5] At the same time, metagenomics is

revolutionizing our view of the tree of life.[6–9] Deep sequencing has proved the existence of previously unimagined bacterial and archaeal lifeforms, inhabiting ever more extreme environments. The genomes of these newly-discovered microorganisms demonstrate there is a stupendous diversity of protein sequences and enzyme functions waiting to be explored.[10–13] Similarly, a combination of culturing and sequencing has massively expanded our knowledge of protist diversity, leading to substantial revisions of the eukaryotic part of the tree.[6,14–16]

The goals of this study were to map our knowledge of enzymes in the context of phylogeny, and thence to draw attention to those parts of the tree of life that remain unexplored by biochemists. Our collective knowledge of enzymes takes many forms, including information on structure, function, kinetic parameters, expression level, folding, stability, mutability, promiscuity and evolution. As a proxy for all of this enzymological knowledge, we have chosen to focus on kinetic parameters for wild type enzymes catalyzing naturally occurring reactions. Previously, Davidi et al. filtered all of the entries in the BRaunschweig ENzyme DAtabase (BRENDA),[17] as it was circa April 2017, for entries of this type. They used this filtered dataset to draw insightful conclusions about the chemical and physiological demands that shape enzyme evolution, while also acknowledging there has been a sampling bias in the enzymes studied to date.[18] Here, we have used their dataset to examine this bias in more detail.

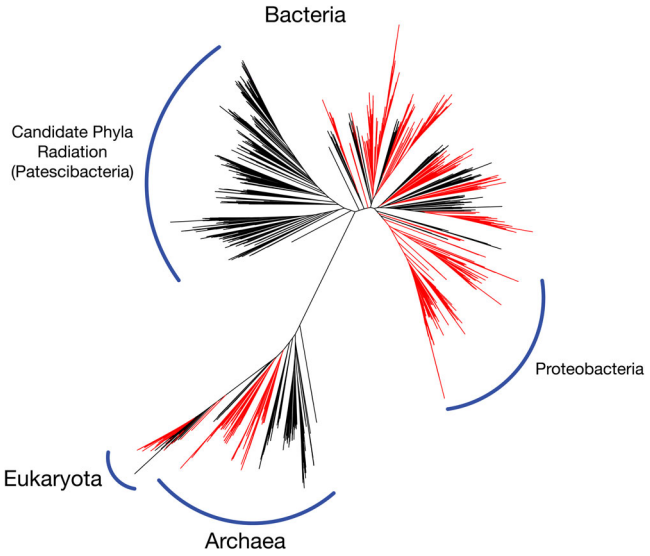

FIGURE 1 Low-resolution mapping of enzymological knowledge onto the tree of life. The phylogenetic tree was constructed by Castelle and Banfield using ribosomal protein sequences[8] and highlights the relative levels of diversity in the three domains of life (Bacteria, Archaea and Eukaryota). We used the Interactive Tree of Life online tool (iTOL)[40] to color red the nodes where there is at least one entry in our BRENDA database of enzyme kinetic parameters. Major parts of the tree with no BRENDA entries are black. The coloration is at nodes approximately corresponding to phyla and aims to emphasize the broad gaps in enzymological knowledge

2 | RESULTS AND DISCUSSION

The dataset compiled by Davidi et al. comprises 12,721 entries (i.e., Table S1 from ref 18, which we accessed in mid-2020). Each entry contains at least one kinetic parameter (k_{cat}, K_M and/or k_{cat}/K_M) for an enzyme, catalyzing either a forward or backward reaction. Usefully, the authors also included the identity of the organism from which the enzyme was obtained. For our taxonomic analyses, we discarded the data on viral enzymes (40 entries), unclassified enzymes (three entries) and an entry on a synthetic ribozyme. This left a trimmed dataset of 12,677 entries. There is inevitably noise in this kind of global analysis, such as occasional inconsistencies between the BRENDA entry and the original paper (as discussed previously[19]). Nevertheless, our dataset serves as an informative snapshot of humankind's effort to study enzymes, as well as where there are gaps in our knowledge.

We began by using the NCBI Taxonomy database[20] to manually assign each BRENDA entry to its domain (Archaea, Bacteria or Eukaryota; Figure 1). In the dataset (Table S1), 901 entries (7%) were from the domain Archaea, 4,804 (38%) were from Bacteria and 6,972 (55%) were from Eukaryota. The eukaryotes comprise the least diverse part of the tree of life,[8,21] and yet – unsurprisingly – over half our accumulated knowledge of enzyme kinetics comes from organisms in this domain. On the other hand, large taxonomic groups such as the bacterial Candidate Phyla Radiation (CPR) and multiple archaeal phyla are unexplored enzymatically (Figure 1).

While Figure 1 provided a low-resolution snapshot across the major phylogenetic groups, we also aimed to rigorously assign each BRENDA entry to one additional taxonomic rank. For Archaea and Bacteria, this was the phylum. There has been considerable recent interest in defining microbial relationships at the level of the phylum,[7,8,22,23] and the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org/) has emerged as a comprehensive, genome-based resource for standardized taxonomic classification.[24] It also serves as a useful standard reference tool for phylogenetically aware enzymology. Therefore, we manually assigned each archaeal and bacterial BRENDA entry to its phylum as defined in release 95 of the GTDB (Table S1, column C). However, some of the GTDB classifications are unlikely to be

familiar to enzymologists, so we also included commonly used identifiers from the NCBI Taxonomy database[20] as an extra column in the dataset (Table S1, column D).

GTDB currently classifies 194,600 genomes into 18 archaeal phyla and 111 bacterial phyla. Only four of the archaeal phyla are represented in our dataset (Figure 2a): Methanobacteriota (354 entries); Thermoproteota (314 entries); Halobacteriota (204 entries); and Thermoplasmatota (26 entries). Similarly, only 24 of 111 bacterial phyla are represented and the data are heavily skewed (Figure 2b). The three most well-characterized phyla are the Proteobacteria (2,403 entries), Firmicutes (748 entries) and Actinobacteriota (722 entries). The Proteobacteria alone account for half of all the bacterial data in BRENDA, while these three phyla combined contribute 81% of the bacterial entries.

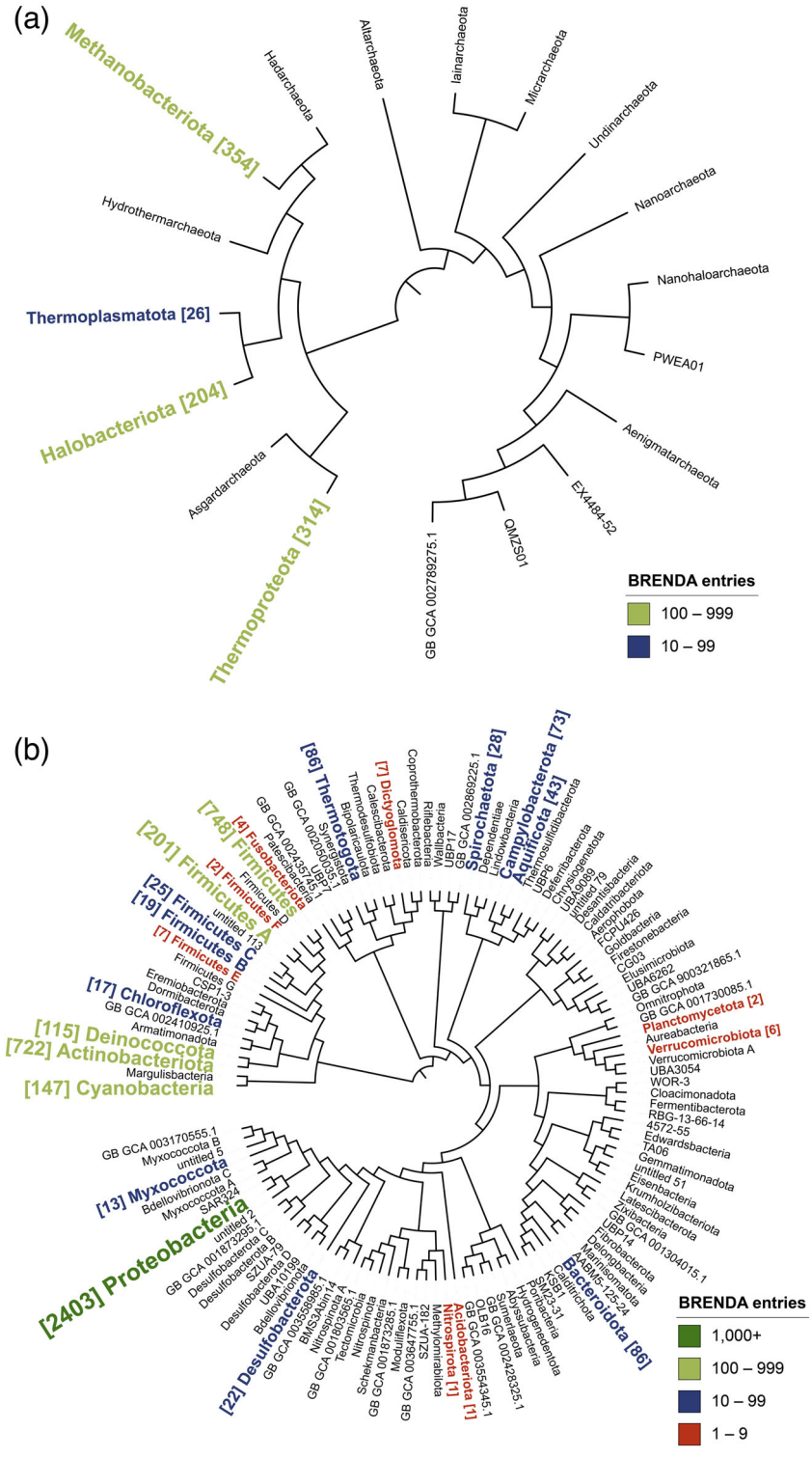

FIGURE 2 Detailed mapping of BRENDA entries onto (a) archaeal and (b) bacterial phyla. The phyla are defined according to the GTDB[24] and the cladograms were from AnnoTree,[41] which we modified using iTOL.[40] Phylum names are color-coded according to the number of BRENDA entries in them, with a black label indicating no entries. The number of BRENDA entries for a phylum (that has at least one) is also shown next to its name on the cladogram

The highest level of eukaryotic relationships is currently best represented by larger-than-kingdom "supergroups".[6] To highlight this level of diversity, we assigned BRENDA entries from eukaryotic species to their supergroup, as defined by Burki et al.[6] (Table S1, column C). In the case of the eukaryotes, the additional information in column D of Table S1 is a lower taxonomic rank. For example, while *H. sapiens* and *S. cerevisiae* are both in the Amorphea supergroup, we have also included their classification in Metazoa (i.e., animals) and Fungi, respectively.

The eukaryotes have most recently been classified into nine supergroups, together with four orphan taxa that remain to be classified.[6] Six of the supergroups and none of the orphan taxa are represented in the enzyme dataset (Figure 3). Our knowledge of enzymology is very heavily biased towards the Amorphea (4,123 entries; 59% of all eukaryote data). Within this supergroup, a few enzymes have been studied from the Amoebozoa but most data come from the Opisthokonta (4,030 entries), which includes animals and fungi. A similar trend was observed for the Archaeplastida supergroup, in which 2,202 of the 2,230 BRENDA entries were from the Chloroplastida (land plants and green algae).

Many major taxonomic groups are notable for their complete absence from, or severe under-representation in, BRENDA. For example, the extremely large group of bacterial lineages known as the Candidate Phyla Radiation (CPR), and classified as the Patescibacteria in GTDB, lack a single entry in the enzyme dataset, even though they constitute a substantial fraction of all the phylogenetic diversity on Earth (Figure 1).[7,8,23] Similarly, the Telonemia-Stramenopila-Alveolata-Rhizaria (TSAR) supergroup is estimated to encompass half of all eukaryotic species diversity[25] and yet it comprises 3% of the eukaryotic BRENDA entries – with almost half of these being on a single alveolate, the malaria parasite *Plasmodium falciparum* (Figure 3 and Table S1).

Finally, we analyzed the species-level diversity in our enzyme dataset. In total, 1,731 species are represented: 78 from Archaea; 703 from Bacteria; and 950 from Eukaryota. Consistent with the bias at higher taxonomic ranks, enzyme data are heavily skewed towards a small number of species (Table 1). Only 18 species have 100 or

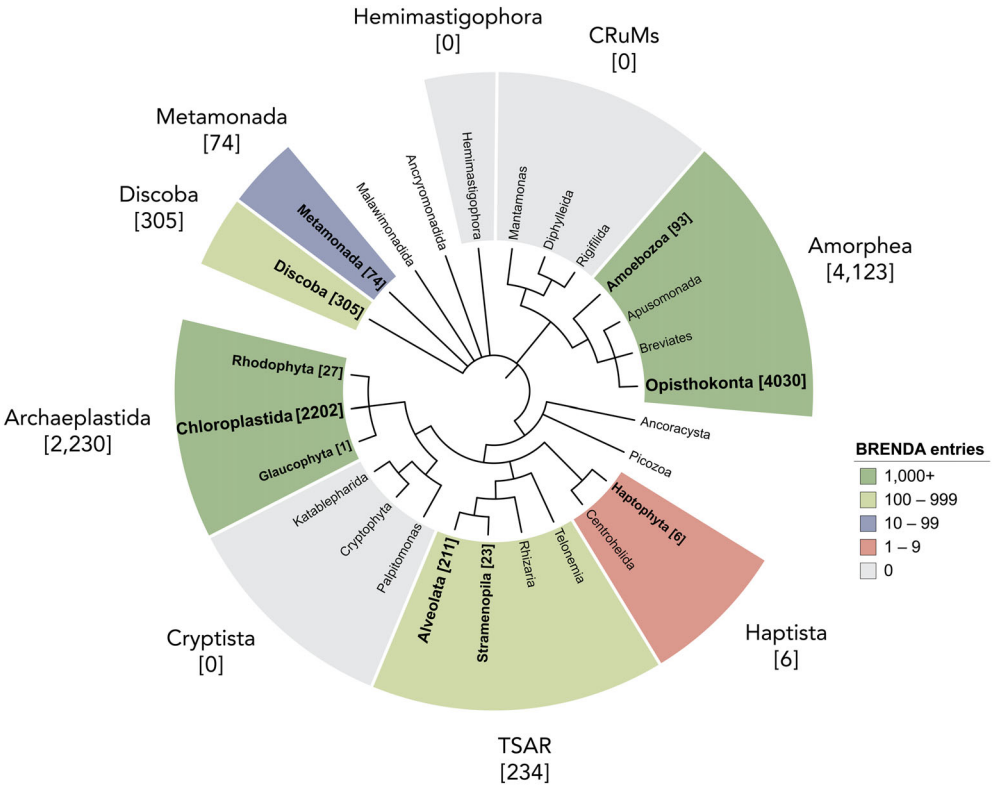

FIGURE 3 Detailed mapping of BRENDA entries onto the major taxonomic groups of eukaryotes, as described by Burki et al.[6] The nine supergroups are shown as sectors that are colored according to their total number of BRENDA entries. Their names, and total number of BRENDA entries, are shown around the outside of the figure. TSAR is the supergroup comprising Telonemia + Stramenopila + Alveolata + Rhizaria. CRuMs is Collodictyonida (synonym Diphylleida) + Rigifilida + Mantamonas. Where there is more than one major clade within a supergroup, these are also shown along with their number of BRENDA entries. The four orphan taxa (not yet assigned to any supergroup) are left unshaded and contribute no entries to BRENDA. The cladogram was constructed with iTOL[40]

TABLE 1 The eight species with the most characterized enzymes, from each domain of life

Domain	No entries	% domain's total entries	Phylum[a] or supergroup[b]	Additional identifier[c]
Archaea				
Sulfolobus solfataricus	107	11.9%	Thermoproteota	Crenarchaeota
Pyrococcus furiosus	69	7.7%	Methanobacteriota	Euryarchaeota
Archaeoglobus fulgidus	67	7.4%	Halobacteriota	Euryarchaeota
Methanocaldococcus jannaschii	57	6.3%	Methanobacteriota	Euryarchaeota
Sulfolobus tokodaii	55	6.1%	Thermoproteota	Crenarchaeota
Pyrococcus horikoshii	46	5.1%	Methanobacteriota	Euryarchaeota
Aeropyrum pernix	38	4.2%	Thermoproteota	Crenarchaeota
Thermococcus kodakarensis	37	4.1%	Methanobacteriota	Euryarchaeota
Bacteria				
Escherichia coli	669	13.9%	Proteobacteria	Gammaproteobacteria
Mycobacterium tuberculosis	195	4.1%	Actinobacteriota	Actinobacteria
Bacillus subtilis	121	2.5%	Firmicutes	Bacilli
Salmonella enterica	112	2.3%	Proteobacteria	Gammaproteobacteria
Pseudomonas putida	107	2.2%	Proteobacteria	Gammaproteobacteria
Pseudomonas aeruginosa	103	2.1%	Proteobacteria	Gammaproteobacteria
Unspecified *Pseudomonas* sp.	101	2.1%	Proteobacteria	Gammaproteobacteria
Thermus thermophilus	84	1.7%	Deinococcota	Deinococcus-Thermus
Eukaryota				
Homo sapiens	766	11.0%	Amorphea	Metazoa (animals)
Rattus norvegicus	548	7.9%	Amorphea	Metazoa (animals)
Saccharomyces cerevisiae	298	4.3%	Amorphea	Fungi
Bos taurus	261	3.7%	Amorphea	Metazoa (animals)
Arabidopsis thaliana	242	3.5%	Archaeplastida	Viridiplantae (green plants)
Mus musculus	207	3.0%	Amorphea	Metazoa (animals)
Sus scrofa	165	2.4%	Amorphea	Metazoa (animals)
Oryctolagus cuniculus	115	1.6%	Amorphea	Metazoa (animals)

[a]For Archaea and Bacteria.
[b]For Eukaryota.
[c]Identifiers from the NCBI Taxonomy browser[20] that may or may not have standing in the taxonomy community, but remain commonly used by enzymologists.

more entries in BRENDA. At the other extreme, 682 species are represented by a single BRENDA entry. Over 88% of the species in our BRENDA dataset are represented by fewer than 10 entries and the trend is consistent across all three domains of life (Figure 4).

The skew towards a few very well-characterized model organisms is most evident in the eukaryotes, where over half of the total entries are from only 2.1% of species and the top eight eukaryotic species contribute 37.3% of all entries (Table 1). The mean number of entries per species is 7.3 but the median is only 2. Across the entire tree of life, seven of the ten most well-characterized species are eukaryotes (Table 1).

Five of the ten most well-characterized species are not only eukaryotes, but mammals (human, rat, cattle, mouse and pig). Together, these five contribute 15% of all BRENDA entries. This observation is unlikely to surprise biochemists, most of whom have probably characterized enzymes from one or more of these species during their undergraduate training. For perspective, it is worth reflecting that there are ~12,000 species of mammals on Earth.[26] However, it has been estimated there are over

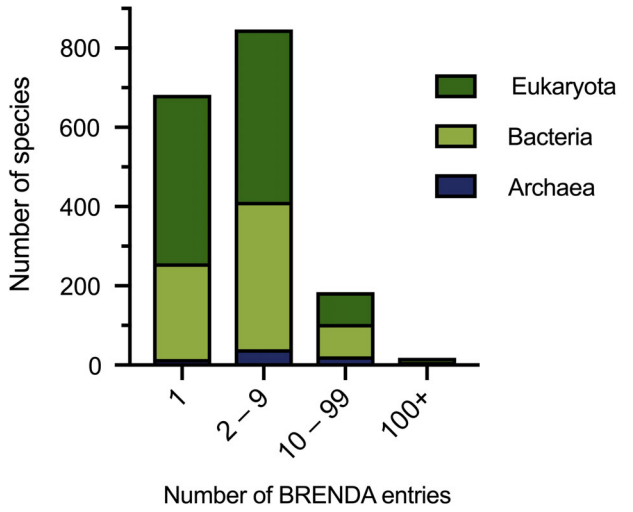

FIGURE 4 Counts of species according to how many entries they contribute to our BRENDA dataset. There are 18 species with more than 100 entries. Of these, one is archaeal, seven are bacterial and ten are eukaryotic

160 million species of animals (almost all arthropods and nematodes), and at least two billion species in total (mostly bacteria).[21] Thus, a conservative estimate is that mammals comprise ~0.0006% of the species on Earth.

Phylogenetic distance increases the likelihood of discovering novel protein folds and functions.[27] This was illustrated in dramatic fashion when a single study reported 1,003 reference genomes selected solely because they maximized coverage of phylogenetic space, and which collectively increased the number of known protein sequence families by over 10%.[11] Sequence similarity networks and genome mining algorithms are also being used increasingly to identify enzymes with divergent sequences that expand our understanding, have biotechnological value, or both.[13,28–32] For example, a recent study identified 33,000 sequences for the carbon-fixing enzyme, ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), in genomic and metagenomic datasets.[28] A subset of 143 maximally diverse enzymes were characterized biochemically. RuBisCO has been studied for decades, from over 200 species, and yet five of these new-to-science enzymes had higher k_{cat} values than any previously characterized. Perhaps surprisingly, the fastest was not from a plant, alga or cyanobacterium, but instead from a soil-dwelling bacterium of the genus *Gallionella*.[28] It remains to be seen whether this fast RuBisCO might overcome the longstanding biotechnological challenge of improving carbon capture, and therefore growth, in crops.[33] There is clearly enormous potential for this approach of phylogenetically aware enzymology to realize the value of the deep sequencing revolution, both by bringing biochemical insights and by uncovering novel biocatalysts with desirable properties for the biotechnology industry.

3 | CONCLUSIONS

Severe taxonomic biases have been quantified in other branches of biology. For example, it is widely acknowledged that the fields of ecology and conservation biology have been overly focused on a handful of charismatic, highly fundable animal species.[26,34–36] In the case of biochemistry, it is equally clear we have a highly biased view of enzymatic diversity at all taxonomic levels. By quantifying that bias here, we have called attention to the vastness of the "enzymatic dark matter" within the tree of life.

In the era of gene synthesis and recombinant protein expression, it is possible to study any enzyme from any organism, regardless of whether that organism is culturable. Studying enzymes from maximally diverse, non-model organisms will broaden our understanding of the sequence-structure–function relationship and shed new light on the metabolic processes taking place throughout the biosphere.[31,37,38] It will also uncover potent natural products and valuable new parts for synthetic biology.[13] Perhaps most importantly, illuminating the enzymatic dark matter will undoubtedly inspire us with the "beautiful aspects of Nature's chemistry".[39]

ACKNOWLEDGEMENTS

This work was supported via a grant from the Marsden Fund to WMP [grant number 18-VUW-050]. DF was supported by the Henry Luce III Fund for Distinguished Scholarship and the James and Jean McClung Endowment for International Research.

AUTHOR CONTRIBUTIONS

Chelsea J. Vickers: Conceptualization; data curation; formal analysis; investigation; visualization; writing-original draft; writing-review and editing. **Dean Fraga:** Conceptualization; data curation; formal analysis; funding acquisition; visualization; writing-review and editing. **Wayne Patrick:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; visualization; writing-original draft; writing-review and editing.

ORCID

Wayne M. Patrick https://orcid.org/0000-0002-2718-8053

REFERENCES

1. Buchner E. Alkoholische Gährung ohne Hefezellen. Ber Dtsch Chem Ges. 1897;30:117–124.

2. Friedmann HC. Alcoholic fermentation without yeast cells. In: A. Cornish-Bowden, editor. New Beer in an Old Bottle: Eduard Buchner and the Growth of Biochemical Knowledge. València, Spain: Universitat de València, 1997; p. 25–31.
3. Heckmann CM, Paradisi F. Looking back: A short history of the discovery of enzymes and how they became powerful chemical tools. Chem Cat Chem. 2020;12:6082–6102.
4. Renata H, Wang ZJ, Arnold FH. Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. Angew Chem Int Ed Engl. 2015;54:3351–3367.
5. Davidi D, Milo R. Lessons on enzyme kinetics from quantitative proteomics. Curr Opin Biotechnol. 2017;46:81–89.
6. Burki F, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. Trends Ecol Evol. 2020;35:43–55.
7. Zhu Q, Mai U, Pfeiffer W, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. Nat Commun. 2019;10:5477.
8. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell. 2018;172:1181–1197.
9. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems. 2018;3:e00055–e00018.
10. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 2018;16:629–645.
11. Mukherjee S, Seshadri R, Varghese NJ, et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol. 2017;35:676–683.
12. León-Zayas R, Peoples L, Biddle JF, et al. The metabolic potential of the single cell genomes obtained from the challenger deep, Mariana trench within the candidate superphylum Parcubacteria (OD1). Environ Microbiol. 2017;19:2769–2784.
13. Gerlt JA. Genomic enzymology: Web tools for leveraging protein family sequence-function space and genome context to discover novel functions. Biochemistry. 2017;56:4293–4308.
14. Gawryluk RMR, Tikhonenkov DV, Hehenberger E, Husnik F, Mylnikov AP, Keeling PJ. Non-photosynthetic predators are sister to red algae. Nature. 2019;572:240–243.
15. Mahé F, de Vargas C, Bass D, et al. Parasites dominate hyperdiverse soil protist communities in neotropical rainforests. Nat Ecol Evol. 2017;1:91.
16. de Vargas C, Audic S, Henry N, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605.
17. Chang A, Jeske L, Ulbrich S, et al. BRENDA, the ELIXIR core data resource in 2021: New developments and updates. Nucleic Acids Res. 2021;49:D498–D508.
18. Davidi D, Longo LM, Jablonska J, Milo R, Tawfik DS. A Bird's-eye view of enzyme evolution: Chemical, physicochemical, and physiological considerations. Chem Rev. 2018;118:8786–8797.
19. Bar-Even A, Noor E, Savir Y, et al. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011;50:4402–4410.
20. Schoch CL, Ciufo S, Domrachev M, et al. NCBI taxonomy: A comprehensive update on curation, resources and tools. Database 2020;2020:baaa062.
21. Larsen BB, Miller EC, Rhodes MK, Wiens JJ. Inordinate fondness multiplied and redistributed: The number of species on earth and the new pie of life. Q Rev Biol. 2017;92:229–265.
22. Baker BJ, De Anda V, Seitz KW, Dombrowski N, Santoro AE, Lloyd KG. Diversity, ecology and evolution of Archaea. Nat Microbiol. 2020;5:887–900.
23. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004.
24. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for bacteria and Archaea. Nat Biotechnol. 2020;38:1079–1086.
25. del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. The others: Our biased perspective of eukaryotic genomes. Trends Ecol Evol. 2014;29:252–259.
26. Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. Taxonomic bias in biodiversity data and societal preferences. Sci Rep. 2017;7:9132.
27. Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA. Myriads of protein families and still counting. Genome Biol. 2003;4:401.
28. Davidi D, Shamshoum M, Guo Z, et al. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. EMBO J. 2020;39:e104081.
29. Hon J, Borko S, Stourac J, et al. EnzymeMiner: Automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res. 2020;48:W104–W109.
30. Mak WS, Tran S, Marcheschi R, et al. Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. Nat Commun. 2015;6:10005.
31. Copp JN, Akiva E, Babbitt PC, Tokuriki N. Revealing unexplored sequence-function space using sequence similarity networks. Biochemistry. 2018;57:4651–4662.
32. Akiva E, Copp JN, Tokuriki N, Babbitt PC. Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily. Proc Natl Acad Sci U S A. 2017;114:E9549–E9558.
33. Ort DR, Merchant SS, Alric J, et al. Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proc Natl Acad Sci U S A. 2015;112:8529–8536.
34. dos Santos JW, Correia RA, Malhado ACM, et al. Drivers of taxonomic bias in conservation research: A global analysis of terrestrial mammals. Anim Conserv. 2020;23:679–688.
35. Clark JA, May RM. Taxonomic bias in conservation research. Science. 2002;297:191–192.
36. Hochkirch A, Samways MJ, Gerlach J, et al. A strategy for the next decade to address data deficiency in neglected biodiversity. Conserv Biol: PMID. 2020;32656858.
37. Ferla MP, Brewster JL, Hall KR, Evans GB, Patrick WM. Primordial-like enzymes from bacteria with reduced genomes. Mol Microbiol. 2017;105:508–524.
38. Newton MS, Arcus VL, Gerth ML, Patrick WM. Enzyme evolution: Innovation is easy optimization is complicated. Curr Opin Struct Biol. 2018;48:110–116.
39. Tawfik DS, van der Donk WA. Editorial overview: Biocatalysis and biotransformation: esoteric niche enzymology. Curr Opin Chem Biol. 2016;31:v–vii.
40. Letunic I, Bork P. Interactive tree of life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 2019;47:W256–W259.
41. Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. AnnoTree: Visualization and exploration of a functionally

annotated microbial tree of life. Nucleic Acids Res. 2019;47: 4442–4448.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Vickers CJ, Fraga D, Patrick WM. Quantifying the taxonomic bias in enzymology. *Protein Science*. 2021;30:914–921. https://doi.org/10.1002/pro.4041