Bacterial amino acid auxotrophies enable energetically costlier proteomes

Niko Kasalo [1], Tomislav Domazet-Lošo [1,2,*] and Mirjana Domazet-Lošo [3,*]

[1]Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia

[2]School of Medicine, Catholic University of Croatia, Ilica 242, HR-10000 Zagreb, Croatia

[3]Department of Applied Computing, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia

*Correspondence: T.D.-L. tdomazet@irb.hr; M. D.-L. mirjana.domazet@fer.hr

Abstract: The outsourcing of amino acid (AA) production to the environment is relatively common across the tree of life. We recently showed that the massive loss of AA synthesis capabilities in animals is governed by selective pressure linked to the energy costs of AA production. Paradoxically, these AA auxotrophies facilitated the evolution of costlier proteomes in animals by enabling the increased use of energetically expensive AAs. Experiments in bacteria have shown that AA auxotrophies can provide a fitness advantage in competition with prototrophic strains. However, it remains unclear whether energy-related selection also drives the evolution of bacterial AA auxotrophies and whether this affects the usage of expensive AAs in bacterial proteomes. To investigate these questions, we computationally determined AA auxotrophy odds across 980 bacterial genomes representing diverse taxa and calculated the energy costs of all their proteins. Here, we show that auxotrophic AAs are generally more expensive to synthesize than prototrophic AAs in bacteria. Moreover, we found that the cost of auxotrophic AAs significantly correlates with the cost of their respective proteomes. Interestingly, out of all considered taxa, Mollicutes and Borreliaceae—chronic pathogens highly successful in immune evasion—have the most AA auxotrophies and code for the most expensive proteomes. These findings indicate that AA auxotrophies in bacteria, similar to those in animals, are shaped by selective pressures related to energy management. Our study reveals that bacterial AA auxotrophies act as costly outsourced functions, enabling bacteria to explore protein sequence space more freely. It

remains to be investigated whether this relaxed use of expensive AAs also enabled auxotrophic bacteria to evolve proteins with improved or novel functionality.

Keywords: bacteria; amino acids; auxotrophy; energy costs; selection; evolution; functional outsourcing

1. Introduction

The concept of functional outsourcing suggests that organisms streamline their genomes by losing costly functions and substituting them through biological interactions [1]. A particularly striking example is the outsourcing of amino acid (AA) production in animals, which almost universally lack the ability to synthesize about half of the proteinogenic AAs [1–6]. Our recent research demonstrates that this phenotype is driven by selective pressures related to the energy cost of AA production, allowing animals to evolve proteomes that incorporate expensive AAs more frequently [2]. These findings raise the possibility that similar correlations between proteome costs and AA biosynthetic capabilities exist in other major clades on the tree of life, including bacteria.

Compared to animals, the understanding of bacterial AA biosynthesis remains relatively incomplete. Earlier studies indicated that most bacteria possess at least a few auxotrophies [7,8]. However, recent research has revealed that the prevalence of AA auxotrophies may have been overestimated due to gaps in knowledge regarding bacterial biosynthetic pathways, though many bacteria still lack the ability to synthesize the full set of AAs [4,9]. Experimental evidence shows that AA auxotrophies can confer a fitness advantage in competition with prototrophic strains [7,10], and they may serve as an evolutionary strategy to reduce biosynthetic burdens via cooperative interactions within microbial communities [11].

However, the adaptive benefits of AA auxotrophies in bacteria remain debated. A previous study speculated that outsourcing AA biosynthesis might reduce cellular metabolic costs [10]. While it is well-established that AAs vary in biosynthetic cost [11–13] and that costlier AAs promote stronger microbial cross-feeding interactions [11], a recent study found no significant correlation between AA biosynthesis costs and the prevalence of AA auxotrophies among bacteria [4].

Nevertheless, alternative approaches could provide a more rigorous test of whether energy-related selection drives the evolution of AA auxotrophies. For instance, if such selection influences the loss of AA biosynthetic capabilities, it should also be reflected in the

incorporation of more expensive AAs into bacterial proteomes [2]. To our knowledge, no study has yet examined bacterial proteome costs in this context.

Detecting auxotrophies remains an active area of research, employing various methodologies. Many studies rely on in silico approaches, such as genome-scale metabolic modeling [14,15] or homology-guided enzyme annotation [4,7,16]. To assess their accuracy, these computational methods are typically validated against a limited number of experimentally derived datasets [4,14]. However, no simple or standardized protocol currently exists for comparing AA auxotrophy estimates across studies.

To address these gaps, we assembled a taxonomically diverse dataset of bacterial proteomes predicted from genome sequences and examined whether AA biosynthesis costs explain trends in AA auxotrophy composition. Using a simple auxotrophy detection methodology based on the MMseqs2 clustering pipeline, we obtained results comparable in quality to previous approaches. Our findings reveal that costlier AAs are more frequently lost and that bacteria with more expensive AA auxotrophies encode costlier proteomes, mirroring patterns observed in animals. These results suggest that energy-driven selection plays a key role in shaping AA auxotrophies in bacteria, enabling them to explore protein sequence space more freely during evolution.

2. Results

2.1. Global trends in bacterial AA auxotrophies

To explore global patterns in the evolution of bacterial auxotrophies, we assembled a database of 980 high-quality proteomes, computationally inferred from their corresponding genomes, capturing broad bacterial diversity (Table S1). We assessed the completeness of AA biosynthesis pathways in these species using an MMseq2-based clustering approach [17]. Our method clusters, in a single step, a representative sample of bacterial enzymes known to catalyze reactions in AA biosynthesis pathways together with all bacterial proteomes in our database (see Materials and Methods). Based on the composition of the recovered clusters, functional information on AA anabolism is then transferred between cluster members, allowing us to determine the completeness of 20 AA biosynthesis pathways for each species. Unlike previous studies that rely on strictly defined cutoff values to designate auxotrophies—thereby losing part of the available information—we analyzed pathway completeness values directly.

This approach provides a more accurate representation of the likelihood that a particular pathway is present.

For each of the twenty amino acids, completeness scores range from 0 to 1, where 0 indicates that all enzymes in the pathway are absent, and 1 indicates that all enzymes are present in a given species (Fig. 1, File S1). When summing the completeness scores for all 20 AAs, we found that the average value across all bacterial species is 17.33, and that nearly 66% of species in our dataset exhibit total pathway completeness above 19. These high completeness values suggest that many bacteria are prototrophic for most amino acids, aligning with findings from recent studies [4,9]. However, certain taxonomic groups—including Lactobacillales, Mollicutes, and Borreliaceae—exhibit substantial incompleteness in multiple amino acid biosynthetic pathways (Fig. 1, File S1), indicating that auxotrophies are common in these groups.

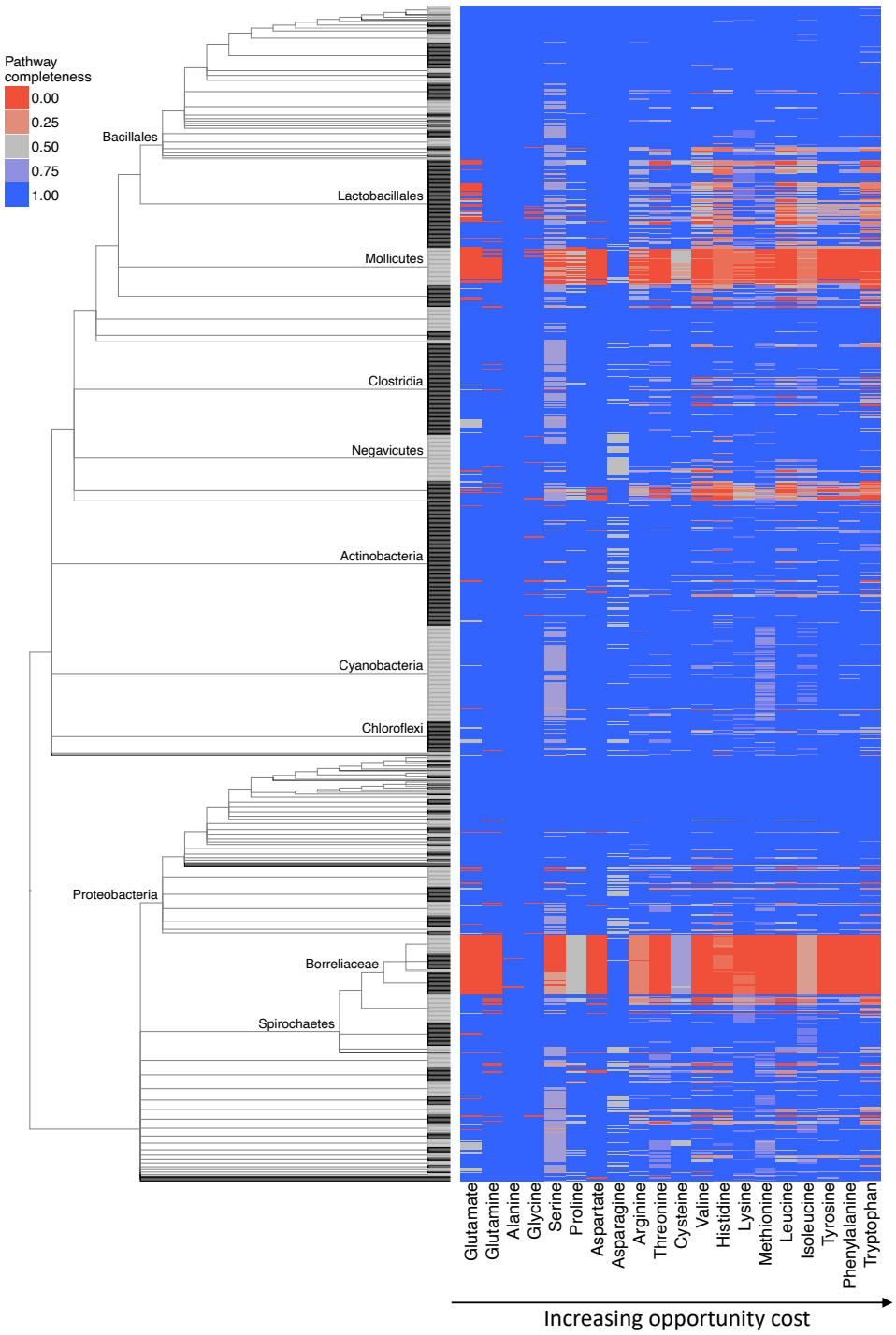

Figure 1. Completeness of AA biosynthesis pathways in bacteria. We created a database of 980 bacterial species to get a comprehensive overview of AA dispensability in this group. Fully resolved tree is shown File S1. We retrieved data on enzymes involved in AA biosynthesis pathways from the KEGG and MetaCyc databases and searched for their homologs within our reference database using MMseqs2 clustering (see Methods). For each AA, we showed a completeness score (CS), which represents the percentage of enzymes within a pathway that returned significant sequence similarity matches to our reference collection of AA biosynthesis

enzymes. In the case of AAs with multiple alternative pathways, we showed the results only for the most complete one.

2.2. Expensive AAs are more commonly lost

If energy-driven selection influences the evolution of bacterial AA auxotrophies, one would expect the ability to synthesize energy-expensive AAs to be lost more frequently. To test this hypothesis, we devised an AA auxotrophy index (AI, Equation 1), defined as 1 minus the completeness score, and compared it against the opportunity cost [2], which estimates the energy expenditure associated with AA synthesis (see Materials and Methods). We observed a significant moderate correlation between opportunity cost and auxotrophy index when opportunity cost values for respiratory metabolism were applied (Fig. 2A, B). However, under fermentative conditions, this correlation was much weaker and not statistically significant (Fig. 2C). Together, these findings suggest that the evolution of AA auxotrophies in bacteria, similar to that in animals, is primarily driven by selection favoring energy savings during AA synthesis—a pattern most pronounced under respiratory conditions.

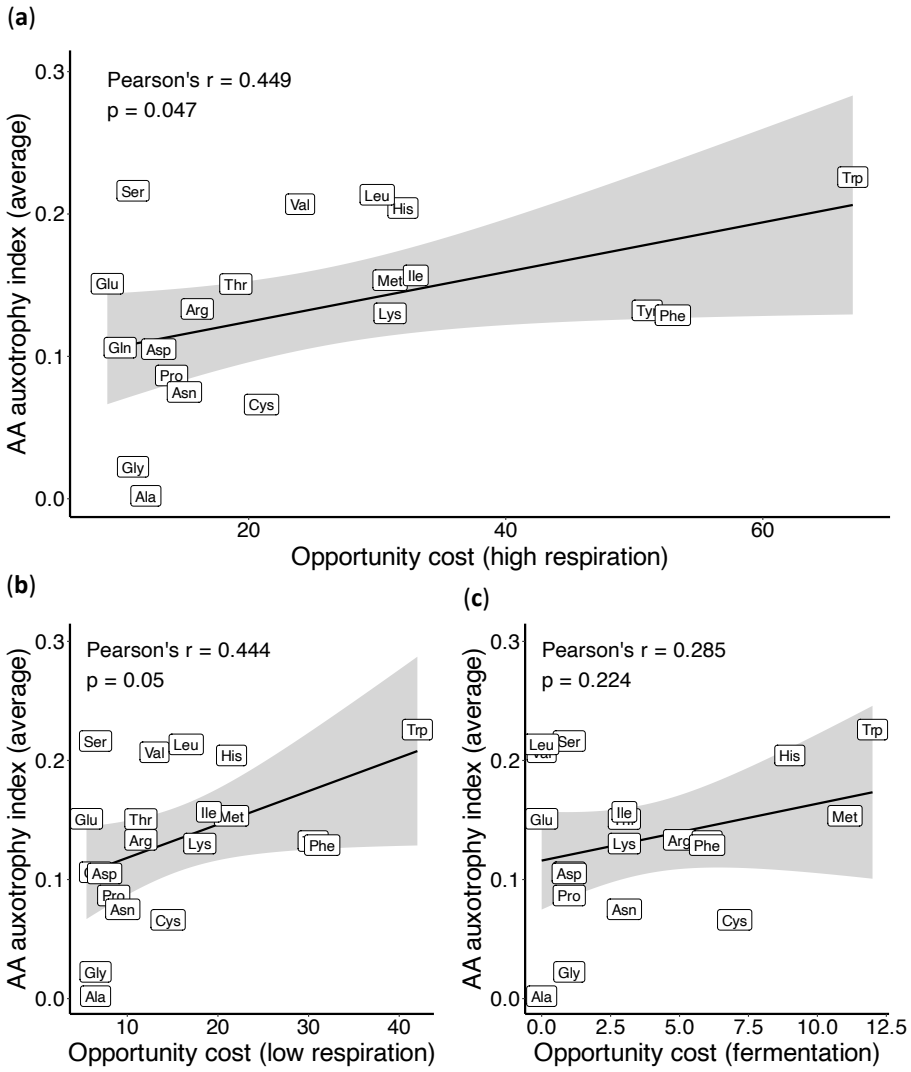

Figure 2. Correlation between AA biosynthesis cost and the average AA auxotrophy index. We estimated the AA auxotrophy index (AI), defined as 1 minus the completeness score (CS), for 980 bacterial species and calculated the average AI value for each AA. This value was then correlated with the opportunity cost of each AA, previously calculated for three different respiratory modes: high respiration, low respiration, and fermentation [2] (see Materials and Methods). The Pearson correlation coefficient and p-value are displayed on the graph.

2.3. Energy savings via AA auxotrophies enable costlier proteomes

However, the central question remains: how do reductions in AA biosynthetic pathways influence the overall cost of encoded proteomes? If energy-driven selection underpins the evolution of AA auxotrophies in bacteria, one would expect auxotrophic species to maintain more expensive proteomes than their prototrophic counterparts. This is because AA auxotrophs

142 expend considerably less energy on AA biosynthesis and acquire the missing AAs at a
143 relatively low cost from the environment. In essence, this pattern suggests that the energy saved
144 on synthesizing costly AAs offsets part of the proteome's energy expenditure, enabling
145 auxotrophs to incorporate more expensive proteins—potentially contributing to novel
146 functions [2].

147 To test this idea, we calculated the opportunity cost of each bacterial proteome ($OC_{proteome}$,
148 Equation 2) as well as the overall biosynthetic savings achieved through AA auxotrophy
149 ($OC_{savings}$, Equation 3). These calculations were performed across three respiratory modes:
150 fermentation, low respiration, and high respiration [2]. We then assessed the correlation
151 between $OC_{proteome}$ and OC_{saving} (Fig. 3). Regardless of the respiratory mode used to estimate
152 energy expenditure, our results showed that species achieving the greatest energy savings in
153 AA production also maintain the most expensive proteomes. Similar to animals, bacterial
154 auxotrophies appear to influence not only the immediate energy budget related to AA
155 production but also reduce selective pressure against the use of energetically costly AAs in
156 proteomes. Notably, we found that Mollicutes and Borreliaceae are almost invariably highly
157 auxotrophic and possess extremely expensive proteomes. On the other hand, Lactobacillales
158 exhibit a wide range of auxotrophy levels and have proteomes that are slightly more expensive
159 than average.

160

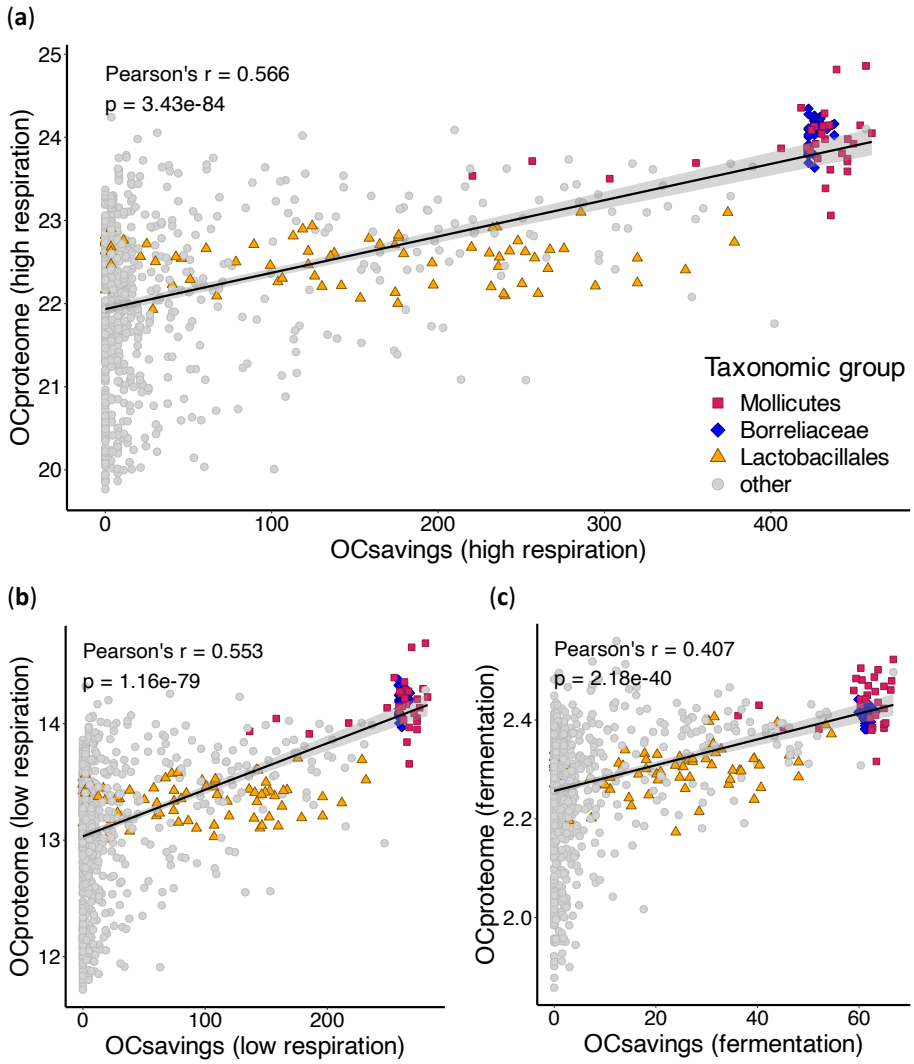

Figure 3. Correlation between the average AA cost per proteome and the amount of outsourced energy for AA biosynthesis. We estimated the completeness of AA biosynthesis pathways for 980 bacterial species. For each proteome, we calculated its opportunity cost ($OC_{proteome}$) by multiplying the opportunity cost of each AA by its frequency in the proteome and summing the resulting values (see Materials and Methods). For each species, we also estimated the energy saved by outsourcing AA biosynthesis ($OC_{savings}$). This was calculated by multiplying the auxotrophy index (AI) by the opportunity cost for each AA and summing the values across all 20 AAs (see Materials and Methods). Opportunity costs for each AA were previously estimated for three different respiratory modes: (**a**) high respiration, (**b**) low respiration, and (**c**) fermentation [2]. The Pearson correlation coefficient (r) and p-value are displayed on the graph.

2.4. Expensive proteins have ecologically relevant functions

To investigate the functional background of the most expensive proteins in Mollicutes, Borreliaceae, and Lactobacillales, we conducted an enrichment analysis of COG functions. Within each of these three clades, we separately calculated the opportunity cost of every protein (Equation 4) and used MMseqs2 clustering [17] to group them into homologous clusters [1,2]. Next, we determined the opportunity cost of each cluster by averaging the opportunity costs of its members (Equation 5). Finally, we performed an enrichment analysis of COG functions for the top 10%, 20%, 30%, 40%, and 50% most expensive clusters within each clade (Fig. 4).

All three bacterial clades show enrichment in a similar set of COG functions; however, the statistical significance of these enrichments varies between the clades. For instance, proteins with unknown functions are among the top two functional categories with the most significant enrichments (i.e., the lowest p-values). However, the second most significantly enriched category differs across clades: defense mechanisms in Borreliaceae (Fig. 4a), inorganic ion transport in Mollicutes (Fig. 4b), and intracellular trafficking in Lactobacillales (Fig. 4c). This suggests that many of the most expensive proteins in these bacterial clades remain severely understudied, while those that have been characterized are involved in different aspects of cellular processes.

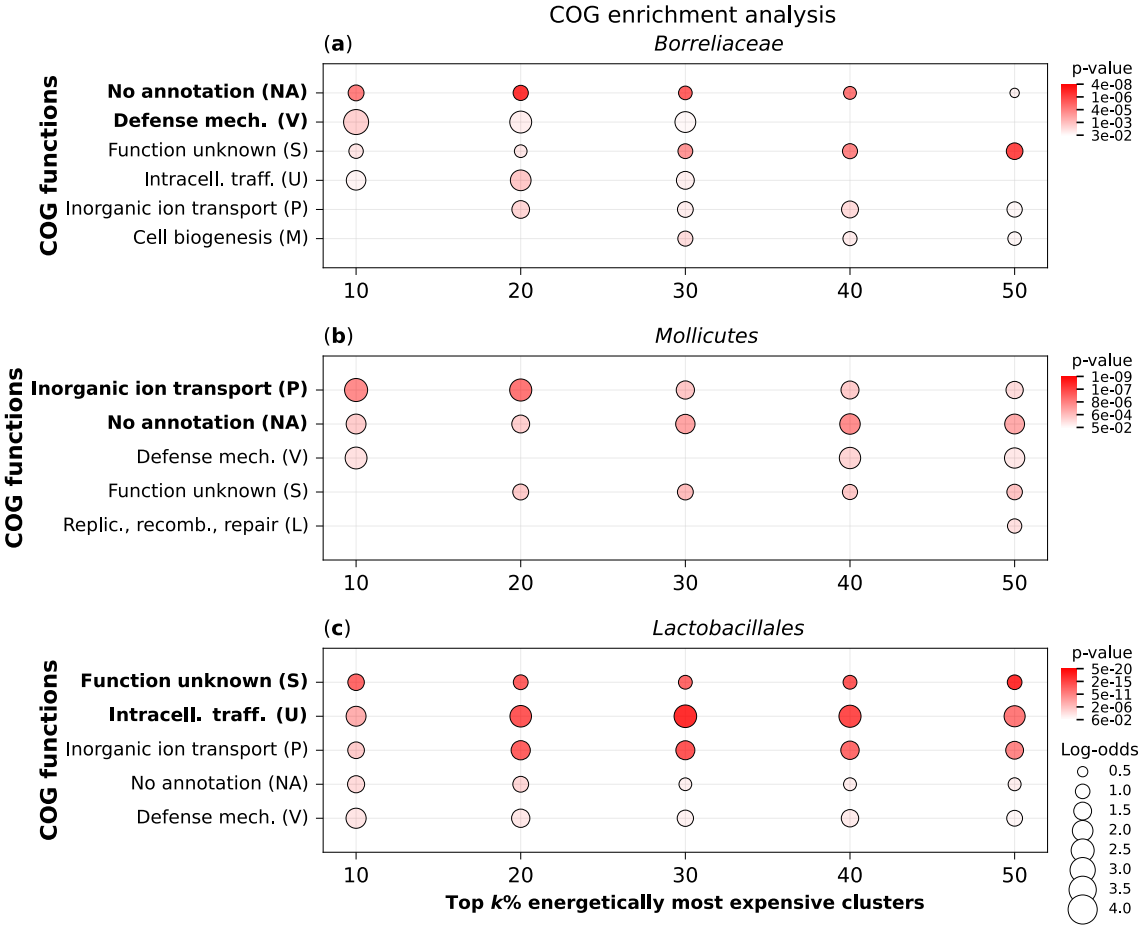

Figure 4. COG Enrichment analysis of the most expensive protein clusters in highly auxotrophic bacterial groups. We analyzed the proteomes of (**a**) Borreliaceae (50 species), (**b**) Mollicutes (31 species), and (**c**) Lactobacillales (74 species) from our full database of 980 bacterial species. Each dataset was clustered separately using the MMseqs2 algorithm to identify clusters of homologous proteins. COG functions were assigned to each cluster using EggNOG-mapper (see Materials and Methods). Clusters without functional annotations were labeled as NA (i.e., no annotation). We performed an overrepresentation analysis for the top k% of clusters ranked by $OC_{cluster}$ (Equation 5), with k = 10, 20, 30, 40, and 50, using a one-tailed hypergeometric test. The resulting p-values were corrected for multiple testing using the Benjamini–Hochberg method. Only enrichment signals with p-values < 0.05 are shown (Table S4).

3. Discussion

204 In our previous work, we applied the concept of functional outsourcing [1] to amino acid (AA)
205 biosynthesis in animals and developed a model describing the conditions required for the
206 evolution of AA auxotrophies [2]. This model predicts that the loss of AA production
207 capabilities is not random because energy-optimizing selection favors the loss of pathways for
208 energetically costly AAs. As a result, AA auxotrophs are able to more freely explore protein
209 sequence space by reducing selective constraints on the use of expensive AAs in their
210 proteomes [2]. To test the broader applicability of this model across major clades, we
211 investigated the patterns of AA auxotrophies in bacteria.

212 While animals exhibit nearly identical sets of auxotrophies among each other, bacterial
213 metabolisms are far more diverse, making bacterial auxotrophies more challenging to detect
214 and interpret [4,7,9]. Although it is known that bacterial AA auxotrophies can confer a fitness
215 advantage in competition with prototrophic strains [7,10], it remains unclear which adaptive
216 benefits are gained through AA auxotrophy and whether they are linked to energy management
217 [2].

218 Our results reveal a significant positive correlation between the AA auxotrophy index averaged
219 across tested bacteria and the AA opportunity cost, as predicted by the AA outsourcing model
220 [2]. This finding aligns with earlier observations that costlier AAs promote stronger microbial
221 cross-feeding interactions [11] and that AA auxotrophies confer a fitness advantage [7,10].
222 Beyond our study, only one prior investigation has explicitly examined bacterial AA
223 auxotrophies from an energy-usage perspective, reporting no significant correlation between
224 the frequency of an AA being auxotrophic and its biosynthetic cost [4]. The reasons for this
225 discrepancy remain unclear and may stem from differences in bacterial genome datasets,
226 auxotrophy detection pipelines, biosynthetic cost estimates, or uneven sampling of bacterial
227 groups. Nonetheless, other trends observed in that study [4] closely resemble our findings—
228 tryptophan, leucine, histidine, valine, and serine are commonly auxotrophic.

229 To further test the robustness of our results, we analyzed data from another comparable study
230 that estimated AA auxotrophies in the gut microbiome using metabolic modeling, though it did
231 not include energy calculations [14]. Similar to our study, this dataset also showed a significant
232 correlation between the frequency of an AA being auxotrophic and its biosynthetic cost, with
233 an even higher correlation coefficient than in our results (Figure S1). Consistent with our
234 findings, they reported that tryptophan, the most expensive AA, is the most commonly
235 auxotrophic. Moreover, their auxotrophy profiles closely resemble ours, despite being derived
236 from a niche-derived community—gut microbiome bacteria [14]. Interestingly, they also found

that host-microbiome and microbe-microbe interactions can play a crucial role in the maintenance and spread of AA auxotrophy [14], aligning with our concept of functional outsourcing [1]. Collectively, these findings suggest that, in at least some bacterial groups, AA auxotrophies are influenced by energy savings at the level of AA biosynthesis.

The second prediction of our model is that bacterial species with more auxotrophies should have more expensive proteomes. Our results confirm this prediction, showing that auxotrophic species indeed maintain genes that encode more expensive proteomes. This finding suggests that AA auxotrophy fundamentally influences protein evolution. The energy savings achieved through outsourcing AA biosynthesis relax the constraints on incorporating costly AAs into proteins, thereby enabling auxotrophic organisms to explore protein sequence space more freely [1,2,18–20].

This is best illustrated by Borreliaceae and Mollicutes, which exhibit extreme levels of both proteome expensiveness and auxotrophy. They have outsourced most of their metabolic processes to their hosts [21–23], making their AAs relatively inexpensive, which likely facilitated pathogen-host coevolution. In contrast, some Lactobacillales species exhibit only a few auxotrophies while others display many, yet this variation does not impact proteome costs, which remain relatively constant. This pattern most likely arises because Lactobacillales predominantly rely on fermentative metabolism [24]. As predicted by our model [2], fermentation lowers AA synthesis costs, thereby relaxing the selective pressures that drive the outsourcing of expensive AA production. This suggests that auxotrophies in Lactobacillales are more strongly determined by the availability of specific AAs in the environment rather than by the energy burden of AA biosynthesis [2].

Based on these findings, we hypothesize that the increase in the frequencies of costly AAs in auxotrophic species could result in the evolution of proteins with novel functions [1,2,18–20]. Interestingly, the most auxotrophic groups with the most expensive proteomes in our analysis are Borreliaceae and Mollicutes, the members of which are notorious for causing severe infections that are difficult to manage [25–27]. It is also known that *Borreliella* (syn. *Borrelia*) *burgdorferi* (Spirochaetales) harbors many Borreliaceae-specific genes with unknown functions, which may be implicated in the development of Lyme disease [28].

Our enrichment analysis further supports the idea that auxotrophies facilitate the evolution of proteins with novel functions. In all three tested groups, many of the most expensive protein clusters have unknown functions, suggesting that the energy saved through outsourcing is

269 invested in lineage-specific proteins. In Borreliaceae, the enrichment of defense-related
270 functions among expensive proteins may be associated with their complex immune evasion
271 strategies [29,30]. Similarly, inorganic ion transport functions are the most expensive category
272 in Mollicutes, whose pathogenicity is closely linked to alterations in the ion transport of host
273 cells [31]. Finally, the enrichment of expensive intracellular trafficking functions in
274 Lactobacillales likely reflects their ability to acquire nutrients through symbiotic interactions
275 [24]. Taken together, our results suggest that auxotrophy-related energy shifts may drive the
276 evolution and function of lineage-specific genes, a topic that warrants further in-depth
277 investigation.

278 In conclusion, our results suggest a global macroevolutionary trend in bacteria where AA
279 biosynthesis capability is shaped by energy-saving selection, ultimately leading to the
280 evolution of more expensive proteomes. However, bacteria exhibit remarkable ecological and
281 metabolic diversity, thriving in a variety of physically and chemically unique habitats
282 [9,32,33]. This diversity raises the possibility that factors beyond energy costs influence
283 reductions in AA biosynthetic pathways. Future studies focusing on specific bacterial lineages
284 and ecologies will be crucial for uncovering the role of such factors in shaping the evolution
285 of AA biosynthetic capabilities.

286

287 **4. Materials and Methods**

288 *2.1. Databases, completeness score and auxotrophy index*

289 To create the database of bacterial proteomes (Table S1), we combined datasets previously
290 assembled for the phylostratigraphic analyses of *Bacillus subtilis* [34] and *Borrelia burgdorferi*
291 [28], along with a resolved phylogeny of *Escherichia coli*. This resulted in a database of 980
292 bacterial proteomes, predicted from their corresponding genomes, representing most major
293 bacterial lineages. The proteomes, primarily retrieved from the NCBI database and
294 supplemented by the Ensembl database, were evaluated for contamination using BUSCO [35],
295 and all were confirmed to be free of contamination [1].

296 In our previous study, we retrieved pathways and enzyme codes involved in AA biosynthesis
297 from the KEGG and MetaCyc databases [2,36]. For AAs that can be synthesized via multiple
298 alternative pathways, we treated each pathway separately, even when they shared some
299 enzymes. Using this collection of enzyme codes associated with AA biosynthesis, we retrieved
300 bacterial protein sequences from the KEGG database [37,38]. For each genus with

representatives annotated in KEGG, we selected the species with the largest number of enzymes catalyzing reactions in AA biosynthesis pathways. This process resulted in a protein sequence reference database comprising 387,892 AA biosynthesis enzymes across 2,095 species (Table S2).

In the next step, we combined the downloaded enzyme sequences known to be involved in AA biosynthesis pathways with all sequences from our bacterial proteomes (4,230,625 sequences across 980 species) into a single database. We then clustered this combined database using MMseqs2 [17] with the following parameters: -cluster-mode 0, -cov-mode 0, -c 0.8, and -e 0.001. Clustering with these parameters generated clusters whose members exhibit highly similar architectures, as the alignment between query and target sequences covered at least 80% of their length [1].

Using the presence of enzymes involved in AA synthesis, we functionally annotated the remaining members of each cluster by transferring functional information from known AA synthesis enzymes. For each AA biosynthesis pathway and species in the database, we calculated a pathway completeness score (CS) by dividing the number of detected enzymes by the total number of enzymes in that pathway, resulting in values ranging from 0 to 1. If a species contained alternative biosynthetic pathways for an AA, the pathway with the highest completeness score was selected. We then calculated the AA auxotrophy index (AI_i) using the following equation:

$$AI_i = 1 - CS_i \tag{1}$$

In this equation CS is completeness score and i denotes one of 20 AAs (i = 1, ..., *20*).

To evaluate the performance of our method, we utilized a previously established testing set of experimentally identified prototrophies and auxotrophies [4,14] to estimate pathway completeness scores (CS). The first tested dataset comprised 160 fully prototrophic species [4]. Our approach exhibited an error rate of 0.012, indicating that approximately 1.2% of amino acids were incorrectly identified as auxotrophic (Data S1). The second dataset included 15 species with at least one known auxotrophy [14]. In this case, we detected an error rate of 0.188 for false prototrophs, meaning that around 19% of amino acids were incorrectly classified as prototrophic (Data S2). Taken together, these error rates suggest that our approach to

auxotrophy detection is conservative and aligns with error rates reported for similar methods [4,14].

2.2. Opportunity cost measures

In our earlier study, we calculated the opportunity cost of biosynthesis for each AA depending on the three respiration modes: high respiration, low respiration and fermentation [2]. The opportunity cost is calculated as the sum of the energy lost in the synthesis of AAs and the energy that would have been produced if a cell catabolized precursors instead of making AAs [2,39,40]. This measure reflects the overall impact of AA synthesis on the cell's energy budget, and is rather consistent regardless of the carbon source used by the bacteria [41]. To approximate how much ATP is generated from the reducing equivalents in different respiratory conditions, we converted the reducing equivalents to ATP in the following way: (i) 'high respiration,' representing fully functional oxidative phosphorylation: 1 NAD(P)H = 2 FADH2 = 2 ATP [12]; (ii) 'low respiration,' representing oxidative phosphorylation without proton pumping at complex I: 1 NAD(P)H = 2 FADH2 = 1 ATP, which corresponds, for instance, to the metabolism of *S. cerevisiae* and some *E. coli* strains [42,43]; (iii) 'fermentation,' representing anaerobic metabolism without conversion of reducing equivalents to ATP [2]. All of these values for each AA are available in Table S3.

Using the AA opportunity cost, we also calculated the opportunity cost of each proteome ($OC_{proteome}$) using the following equation:

$$OC_{proteome} = \frac{\sum_{i=1}^{n=20} OC_i \times N_i}{\sum_{i=1}^{n=20} N_i} = \sum_{i=1}^{n=20} OC_i \times F_i \quad (2)$$

In this equation, $OC_{proteome}$ is a weighted mean where OC_i represents the opportunity cost of the i-th AA, N_i denotes the total number of occurrences of this AA in the entire proteome, and F_i represents the frequency of the AA in the proteome (calculated as the number of occurrences of the AA divided by the total number of AAs in the proteome).

We also introduced a new measure ($OC_{savings}$) to quantify the energy savings of a species by linking the AA auxotrophy index (AI_i) to the opportunity cost. This measure estimates the energy saved by outsourcing AA production to the environment. It is calculated as follows:

362 $$OC_{savings} = \sum_{i=1}^{n=20} OC_i \times AI_i \quad (3)$$

363

364 In this equation, OC_i denotes the opportunity cost of a given AA, while AI_i denotes the AA
365 auxotrophy index.

366 *2.2. COG functions enrichment analyses*

367 For functional analyses, we analyzed separately the proteomes of Borreliaceae (50), Mollicutes
368 (31), and Lactobacillales (74) taken from our full database of 980 bacterial species. We
369 clustered the three datasets separately using the MMseqs2 cluster algorithm (the 14-7e284
370 version) with the following parameters: -e 0.001 -c 0.8 --max-seqs 400 --cluster-mode 1 [1,17].
371 For each protein in the datasets we obtained its COG annotations using the EggNOG-mapper
372 (version 2.1.12) [44] with the diamond (version 2.1.8) searching tool [45]. We also computed
373 for each protein its opportunity cost ($OC_{protein}$) for high respiration mode [2] as:

374

375 $$OC_{protein} = \frac{\sum_{i=1}^{n=20} OC_i \times n_i}{\sum_{i=1}^{n=20} n_i} = \sum_{i=1}^{n=20} OC_i \times f_i \quad (4)$$

376

377 In this equation, $OC_{protein}$ is a weighted mean where OC_i denotes the opportunity cost of the *i*-
378 th amino acid, n_i is the number of occurrences of the *i*-th of amino acid in a protein and f_i is the
379 frequency of the *i*-th amino acid in a protein.

380 Finally, we performed the functional enrichment analysis of the three datasets independently.
381 For each dataset, a cluster was assigned a COG function if at least one of its members was
382 annotated with that function. The clusters whose proteins had no annotations were assigned
383 with NA. The enrichment analysis was done for clusters with at least 10 members. For each
384 cluster we calculated the average opportunity cost of a cluster ($OC_{cluster}$) as follows:

385

386 $$OC_{cluster} = \frac{\sum_{i=1}^{N} OC_{protein(i)}}{N_{cluster}} \quad (5)$$

387

388 In this equation, $OC_{protein(i)}$ is the opportunity cost of the *i*-th protein in a cluster (see Equation
389 4) and $N_{cluster}$ is the number of proteins in the cluster.

We performed an overrepresentation analysis for the top k% of clusters ranked by $OC_{cluster}$, with $k = 10, 20, 30, 40, 50$, using a one-tailed hypergeometric test as implemented in the Python *scipy.stats* module. The obtained p-values were corrected for multiple testing and adjusted using the Benjamini–Hochberg method as implemented in the Python *statsmodels* library [46]. All results of enrichment analysis are shown in Table S4.

To calculate correlations, we used the cor.test() function in the R stats (v. 3.6.2) package. The heatmap was visualized using the ggtree R package [47].

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Table S1: Bacterial database; Table S2: Enzyme reference database; Table S3: Amino acid biosynthesis costs; Table S4: Results of enrichment analysis; Data S1: Prototrophy evaluation; Data S2: Auxotrophy evaluation; File S1: Fully resolved tree with AA auxotrophy heatmap; Figure S1: Correlation between AA biosynthesis cost and the average AA auxotrophy index for Starke et al. 2023 data.

Author Contributions: Conceptualization, N.K., T.D.-L. and M.D.-L.; methodology, N.K., T.D.-L. and M.D.-L.; software, N.K. and M.D.-L.; validation, N.K., T.D.-L. and M.D.-L.; formal analysis, N.K., T.D.-L. and M.D.-L.; writing—original draft preparation, N.K., T.D.-L. and M.D.-L.; writing—review and editing, N.K., T.D.-L. and M.D.-L.; visualization, N.K.; supervision, T.D.-L. and M.D.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Regional Development Fund KK.01.1.1.01.0009 DATACROSS (M.D.-L., T.D.-L.).

Data Availability Statement: All data are available in the supplementary materials.

Acknowledgments: We thank M. Futo, A. Tušar, S. Koska, D. Franjević and G. Klobučar for discussions. We used the computational resources of the University Computing Center - SRCE (Padobran) and the Institute Ruđer Bošković.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Domazet-Lošo, M.; Široki, T.; Šimičević, K.; Domazet-Lošo, T. Macroevolutionary Dynamics of Gene Family Gain and Loss along Multicellular Eukaryotic Lineages. *Nat Commun* **2024**, *15*, 2663, doi:10.1038/s41467-024-47017-w.

2. Kasalo, N.; Domazet-Lošo, M.; Domazet-Lošo, T. Massive Outsourcing of Energetically Costly Amino Acids at the Origin of Animals. *bioRxiv* 2024, https://doi.org/10.1101/2024.04.18.590100.

3. Payne, S.H.; Loomis, W.F. Retention and Loss of Amino Acid Biosynthetic Pathways Based on Analysis of Whole-Genome Sequences. *Eukaryotic Cell* **2006**, *5*, 272–276, doi:10.1128/ec.5.2.272-276.2006.

4. Ramoneda, J.; Jensen, T.B.N.; Price, M.N.; Casamayor, E.O.; Fierer, N. Taxonomic and Environmental Distribution of Bacterial Amino Acid Auxotrophies. *Nat Commun* **2023**, *14*, 7608, doi:10.1038/s41467-023-43435-4.

5. Richter, D.J.; Fozouni, P.; Eisen, M.B.; King, N. Gene Family Innovation, Conservation and Loss on the Animal Stem Lineage. *eLife* **2018**, *7*, e34226, doi:10.7554/eLife.34226.

6. Guedes, R.; Prosdocimi, F.; Fernandes, G.; Moura, L.; Ribeiro, H.; Ortega, J. Amino Acids Biosynthesis and Nitrogen Assimilation Pathways: A Great Genomic Deletion during Eukaryotes Evolution. *BMC Genomics* **2011**, *12*, S2, doi:10.1186/1471-2164-12-S4-S2.

7. D'Souza, G.; Waschina, S.; Pande, S.; Bohl, K.; Kaleta, C.; Kost, C. Less Is More: Selective Advantages CAN Explain the Prevalent Loss of Biosynthetic Genes in Bacteria. *Evolution* **2014**, *68*, 2559–2570, doi:10.1111/evo.12468.

8. Mee, M.T.; Wang, H.H. Engineering Ecosystems and Synthetic Ecologies. *Mol. BioSyst.* **2012**, *8*, 2470, doi:10.1039/c2mb25133g.

9. Price, M.N.; Zane, G.M.; Kuehl, J.V.; Melnyk, R.A.; Wall, J.D.; Deutschbauer, A.M.; Arkin, A.P. Filling Gaps in Bacterial Amino Acid Biosynthesis Pathways with High-Throughput Genetics. *PLoS Genet* **2018**, *14*, e1007147, doi:10.1371/journal.pgen.1007147.

10. D'Souza, G.; Kost, C. Experimental Evolution of Metabolic Dependency in Bacteria. *PLoS Genet* **2016**, *12*, e1006364, doi:10.1371/journal.pgen.1006364.

11. Mee, M.T.; Collins, J.J.; Church, G.M.; Wang, H.H. Syntrophic Exchange in Synthetic Microbial Communities. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, doi:10.1073/pnas.1405641111.

12. Kaleta, C.; Schäuble, S.; Rinas, U.; Schuster, S. Metabolic Costs of Amino Acid and Protein Production in *Escherichia Coli*. *Biotechnology Journal* **2013**, *8*, 1105–1114, doi:10.1002/biot.201200267.

13. Zengler, K.; Zaramela, L.S. The Social Network of Microorganisms — How Auxotrophies Shape Complex Communities. *Nat Rev Microbiol* **2018**, *16*, 383–390, doi:10.1038/s41579-018-0004-5.

14. Starke, S.; Harris, D.M.M.; Zimmermann, J.; Schuchardt, S.; Oumari, M.; Frank, D.; Bang, C.; Rosenstiel, P.; Schreiber, S.; Frey, N.; et al. Amino Acid Auxotrophies in Human Gut Bacteria Are Linked to Higher Microbiome Diversity and Long-Term Stability. *The ISME Journal* **2023**, *17*, 2370–2380, doi:10.1038/s41396-023-01537-3.

15. Zimmermann, J.; Kaleta, C.; Waschina, S. Gapseq: Informed Prediction of Bacterial Metabolic Pathways and Reconstruction of Accurate Metabolic Models. *Genome Biol* **2021**, *22*, 81, doi:10.1186/s13059-021-02295-1.

16. Price, M.N.; Deutschbauer, A.M.; Arkin, A.P. GapMind: Automated Annotation of Amino Acid Biosynthesis. *mSystems* **2020**, *5*, 10.1128/msystems.00291-20, doi:10.1128/msystems.00291-20.

17. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, *35*, 1026–1028, doi:10.1038/nbt.3988.

18. Domazet-Lošo, T.; Brajković, J.; Tautz, D. A Phylostratigraphy Approach to Uncover the Genomic History of Major Adaptations in Metazoan Lineages. *Trends in Genetics* **2007**, *23*, 533–539, doi:10.1016/j.tig.2007.08.014.

19. Domazet-Loso, T.; Tautz, D. An Evolutionary Analysis of Orphan Genes in Drosophila. *Genome Res.* **2003**, *13*, 2213–2219, doi:10.1101/gr.1311003.

20. Tautz, D.; Domazet-Lošo, T. The Evolutionary Origin of Orphan Genes. *Nat Rev Genet* **2011**, *12*, 692–702, doi:10.1038/nrg3053.

21. Kerstholt, M.; Netea, M.G.; Joosten, L.A.B. Borrelia Burgdorferi Hijacks Cellular Metabolism of Immune Cells: Consequences for Host Defense. *Ticks and Tick-borne Diseases* **2020**, *11*, 101386, doi:10.1016/j.ttbdis.2020.101386.

22. Sirand-Pugnet, P.; Citti, C.; Barré, A.; Blanchard, A. Evolution of Mollicutes: Down a Bumpy Road with Twists and Turns. *Research in Microbiology* **2007**, *158*, 754–766, doi:10.1016/j.resmic.2007.09.007.

23. Fisunov, G.Y.; Alexeev, D.G.; Bazaleev, N.A.; Ladygina, V.G.; Galyamina, M.A.; Kondratov, I.G.; Zhukova, N.A.; Serebryakova, M.V.; Demina, I.A.; Govorun, V.M. Core

481 Proteome of the Minimal Cell: Comparative Proteomics of Three Mollicute Species. *PLoS*
482 *ONE* **2011**, *6*, e21964, doi:10.1371/journal.pone.0021964.

483 24. Makarova, K.; Slesarev, A.; Wolf, Y.; Sorokin, A.; Mirkin, B.; Koonin, E.; Pavlov, A.;
484 Pavlova, N.; Karamychev, V.; Polouchine, N.; et al. Comparative Genomics of the Lactic Acid
485 Bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15611–15616,
486 doi:10.1073/pnas.0607117103.

487 25. Pilo, P.; Frey, J.; Vilei, E.M. Molecular Mechanisms of Pathogenicity of Mycoplasma
488 Mycoides Subsp. Mycoides SC. *The Veterinary Journal* **2007**, *174*, 513–521,
489 doi:10.1016/j.tvjl.2006.10.016.

490 26. Strnad, M.; Rudenko, N.; Rego, R.O.M. Pathogenicity and Virulence of *Borrelia*
491 *Burgdorferi*. *Virulence* **2023**, *14*, 2265015, doi:10.1080/21505594.2023.2265015.

492 27. Waites, K.B.; Talkington, D.F. *Mycoplasma Pneumoniae* and Its Role as a Human
493 Pathogen. *Clin Microbiol Rev* **2004**, *17*, 697–728, doi:10.1128/CMR.17.4.697-728.2004.

494 28. Čorak, N.; Anniko, S.; Daschkin-Steinborn, C.; Krey, V.; Koska, S.; Futo, M.; Široki, T.;
495 Woichansky, I.; Opašić, L.; Kifer, D.; et al. Pleomorphic Variants of Borreliella (Syn. Borrelia)
496 Burgdorferi Express Evolutionary Distinct Transcriptomes. *IJMS* **2023**, *24*, 5594,
497 doi:10.3390/ijms24065594.

498 29. Anderson, C.; Brissette, C.A. The Brilliance of Borrelia: Mechanisms of Host Immune
499 Evasion by Lyme Disease-Causing Spirochetes. *Pathogens* **2021**, *10*, 281,
500 doi:10.3390/pathogens10030281.

501 30. Dulipati, V.; Meri, S.; Panelius, J. Complement Evasion Strategies of *Borrelia Burgdorferi*
502 Sensu Lato. *FEBS Letters* **2020**, *594*, 2645–2656, doi:10.1002/1873-3468.13894.

503 31. Lambert, L.C.; Trummell, H.Q.; Singh, A.; Cassell, G.H.; Bridges, R.J. *Mycoplasma*
504 *Pulmonis* Inhibits Electrogenic Ion Transport across Murine Tracheal Epithelial Cell
505 Monolayers. *Infect Immun* **1998**, *66*, 272–279, doi:10.1128/IAI.66.1.272-279.1998.

506 32. Flemming, H.-C.; Wuertz, S. Bacteria and Archaea on Earth and Their Abundance in
507 Biofilms. *Nat Rev Microbiol* **2019**, *17*, 247–260, doi:10.1038/s41579-019-0158-9.

508 33. Lozupone, C.A.; Knight, R. Global Patterns in Bacterial Diversity. *Proc. Natl. Acad. Sci.*
509 *U.S.A.* **2007**, *104*, 11436–11440, doi:10.1073/pnas.0611525104.

34. Futo, M.; Opašić, L.; Koska, S.; Čorak, N.; Široki, T.; Ravikumar, V.; Thorsell, A.; Lenuzzi, M.; Kifer, D.; Domazet-Lošo, M.; et al. Embryo-Like Features in Developing *Bacillus Subtilis* Biofilms. *Molecular Biology and Evolution* **2021**, *38*, 31–47, doi:10.1093/molbev/msaa217.

35. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.

36. Caspi, R.; Billington, R.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Ong, W.K.; Paley, S.; Subhraveti, P.; Karp, P.D. The MetaCyc Database of Metabolic Pathways and Enzymes - a 2019 Update. *Nucleic Acids Research* **2020**, *48*, D445–D453, doi:10.1093/nar/gkz862.

37. Kanehisa, M.; Furumichi, M.; Sato, Y.; Kawashima, M.; Ishiguro-Watanabe, M. KEGG for Taxonomy-Based Analysis of Pathways and Genomes. *Nucleic Acids Research* **2023**, *51*, D587–D592, doi:10.1093/nar/gkac963.

38. Kanehisa, M. The KEGG Database. In *Novartis Foundation Symposia*; Bock, G., Goode, J.A., Eds.; Wiley, 2002; Vol. 247, pp. 91–103 ISBN 978-0-470-84480-9.

39. Craig, C.L.; Weber, R.S. Selection Costs of Amino Acid Substitutions in ColE1 and ColIa Gene Clusters Harbored by Escherichia Coli. *Molecular Biology and Evolution* **1998**, *15*, 774–776, doi:10.1093/oxfordjournals.molbev.a025981.

40. Zhang, H.; Wang, Y.; Li, J.; Chen, H.; He, X.; Zhang, H.; Liang, H.; Lu, J. Biosynthetic Energy Cost for Amino Acids Decreases in Cancer Evolution. *Nat Commun* **2018**, *9*, 4124, doi:10.1038/s41467-018-06461-1.

41. Akashi, H.; Gojobori, T. Metabolic Efficiency and Amino Acid Composition in the Proteomes of Escherichia Coli and Bacillus Subtilis. *Proceedings of the National Academy of Sciences* **2002**, *99*, 3695–3700, doi:10.1073/pnas.062526999.

42. Kok, S.; Kozak, B.U.; Pronk, J.T.; Maris, A.J.A. Energy Coupling in Saccharomyces Cerevisiae: Selected Opportunities for Metabolic Engineering. *FEMS Yeast Res* **2012**, *12*, 387–397, doi:10.1111/j.1567-1364.2012.00799.x.

43. Schuetz, R.; Kuepfer, L.; Sauer, U. Systematic Evaluation of Objective Functions for Predicting Intracellular Fluxes in *Escherichia Coli*. *Molecular Systems Biology* **2007**, *3*, 119, doi:10.1038/msb4100162.

541 44. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.;
542 Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A Hierarchical,
543 Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms
544 and 2502 Viruses. *Nucleic Acids Research* **2019**, *47*, D309–D314, doi:10.1093/nar/gky1085.

545 45. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using
546 DIAMOND. *Nat Methods* **2015**, *12*, 59–60, doi:10.1038/nmeth.3176.

547 46. Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with
548 Python.; Austin, Texas, 2010; pp. 92–96.

549 47. Yu, G.; Smith, D.K.; Zhu, H.; Guan, Y.; Lam, T.T. Ggtree: An Package for
550 Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other
551 Associated Data. *Methods Ecol Evol* **2017**, *8*, 28–36, doi:10.1111/2041-210X.12628.

552

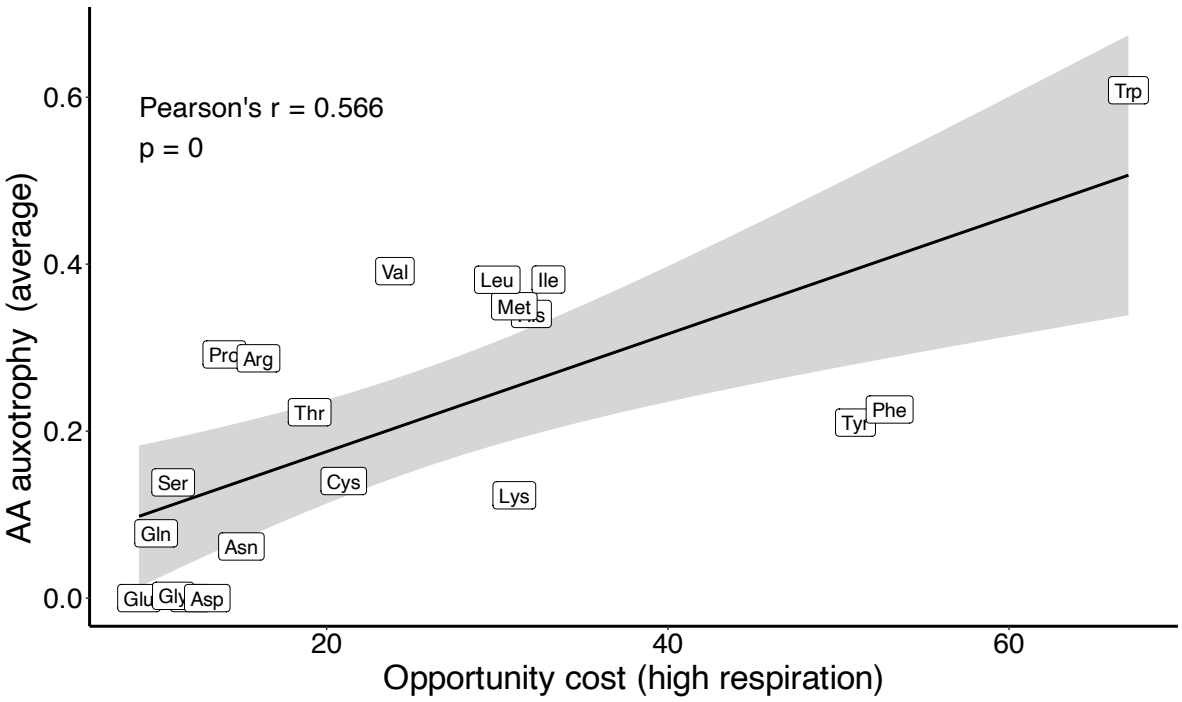

Figure S1. Correlation between AA biosynthesis cost and the average AA auxotrophy based on data from Starke et al. (2023). We calculated the average of AA auxotrophy measures provided in the Starke et al. (2023) study for 3687 bacterial species. We correlated this value with the opportunity cost of each AA, calculated for high respiration mode (see Materials and Methods). Pearson correlation coefficient and p-value are shown on the graph.