PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Physicochemical classification of organisms

Eloy Vallina Estrada[a] and Mikael Oliveberg[a,1]

Edited by Martin Gruebele, University of Illinois at Urbana–Champaign, Urbana, IL; received December 20, 2021; accepted March 16, 2022

The hypervariable residues that compose the major part of proteins' surfaces are generally considered outside evolutionary control. Yet, these "nonconserved" residues determine the outcome of stochastic encounters in crowded cells. It has recently become apparent that these encounters are not as random as one might imagine, but carefully orchestrated by the intracellular electrostatics to optimize protein diffusion, interactivity, and partner search. The most influential factor here is the protein surface-charge density, which takes different optimal values across organisms with different intracellular conditions. In this study, we examine how far the net-charge density and other physicochemical properties of proteomes will take us in terms of distinguishing organisms in general. The results show that these global proteome properties not only follow the established taxonomical hierarchy, but also provide clues to functional adaptation. In many cases, the proteome–property divergence is even resolved at species level. Accordingly, the variable parts of the genes are not as free to drift as they seem in sequence alignment, but present a complementary tool for functional, taxonomic, and evolutionary assignment.

proteome properties | taxonomy | protein electrostatics | intracellular diffusion | functional evolution

Significance

A common way to establish functional relations across organisms is through sequence conservation. In this study, we show that functional evolution also involves a second level of optimization coded in sequence regions that are generally considered nonconserved: the charge of the hypervariable surfaces of folded proteins. The physicochemical signatures of this protein-surface optimization not only diverge in a phylogenetically consistent way, but also are so systematically organized that they allow for detailed functional classification of organisms across all kingdoms of life. Accordingly, the finding 1) exposes a universal determinant of cellular fitness and 2) presents an alternative method for evolutionary and functional elucidation of organisms that is orthogonal to conventional sequence alignment.

Recent studies of live cells reveal that cytosolic crowding imposes some unique functional challenges that have previously been unconsidered. Essentially, the cytosolic proteins are not just sterically obstructive, but also interact electrostatically with one another through repulsive and attractive forces (1). These diffusive interactions are commonly referred to as "quinary interactions" (2), and their effect on the proteins largely exceeds that predicted from simplistic hard-sphere crowding models (3). The most dramatic effect of altering the quinary interactions is observed for the protein motions. A protein that normally diffuses relatively freely in the cytosolic compartment can be changed to get stuck to the intracellular surrounding by a single surface mutation (4). The principal determinant behind this effect is the protein-surface charge (4, 5). To maintain the cytosolic components suitably fluid, most biological macromolecules, like proteins, nucleic acids, and membranes, carry repulsive net-negative charge, and complete loss of this repulsion will naturally promote aggregation and functional arrest (1). However, the role of the protein charge has turned out to be more subtle than that: It modulates in detail the functional protein–protein encounters (1, 5–8). Because the strength and duration of these dynamic encounters need to be kept within certain limits for the cell to function optimally, the protein-charge decoration itself has been suggested to be under biological control (1, 4, 9–11). This very idea challenges the notion that the composition of the variable protein surfaces drifts freely and adds another dimension to protein evolution and organism fitness (12). Attention then shifts from the relatively small and highly conserved binding interfaces and active sites visible in crystal structures to the least-conserved parts of the protein surfaces exposed to the cytosolic surrounding. Proteome-wide studies of *Escherichia coli* confirm that there is indeed a systematic bias toward negative charge density and show also that not any negative charge density is acceptable: Proteins distribute around a moderately negative value, away from which few deviations are observed (1, 13) (Fig. 1). Similar results are obtained from measurements of isoelectric points, leading to the conclusion that the majority of soluble proteins are acidic and that the degree of this acidity varies across organisms (14–17). Classical examples are the proteomes of some halophilic archaea, with net-charge densities 10 times more negative than observed for most other organisms (18–23). Together, these findings show that the variable protein surfaces contain previously unrecognized evolutionary cues, which can be captured in terms of specific sets of physicochemical properties. The question is then whether organism identity can be deduced from physicochemical observables alone. To explore this possibility, we map here the divergence of proteome properties across organisms against the established taxonomic classification and demonstrate that the resolving power is indeed remarkably high. The results show that distinct clustering and separations of proteome properties not only follow taxonomic divisions, but also reflect their adaptation to

Author affiliations: [a]Department of Biochemistry and Biophysics, Arrhenius Laboratories of Natural Sciences, Stockholm University, S-106 91 Stockholm, Sweden

Author contributions: E.V.E. and M.O. designed research; E.V.E. and M.O. performed research; E.V.E. contributed new reagents/analytic tools; E.V.E. and M.O. analyzed data; and E.V.E. and M.O. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mikael.oliveberg@dbb.su.se.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2122957119/-/DCSupplemental.

Published May 2, 2022.

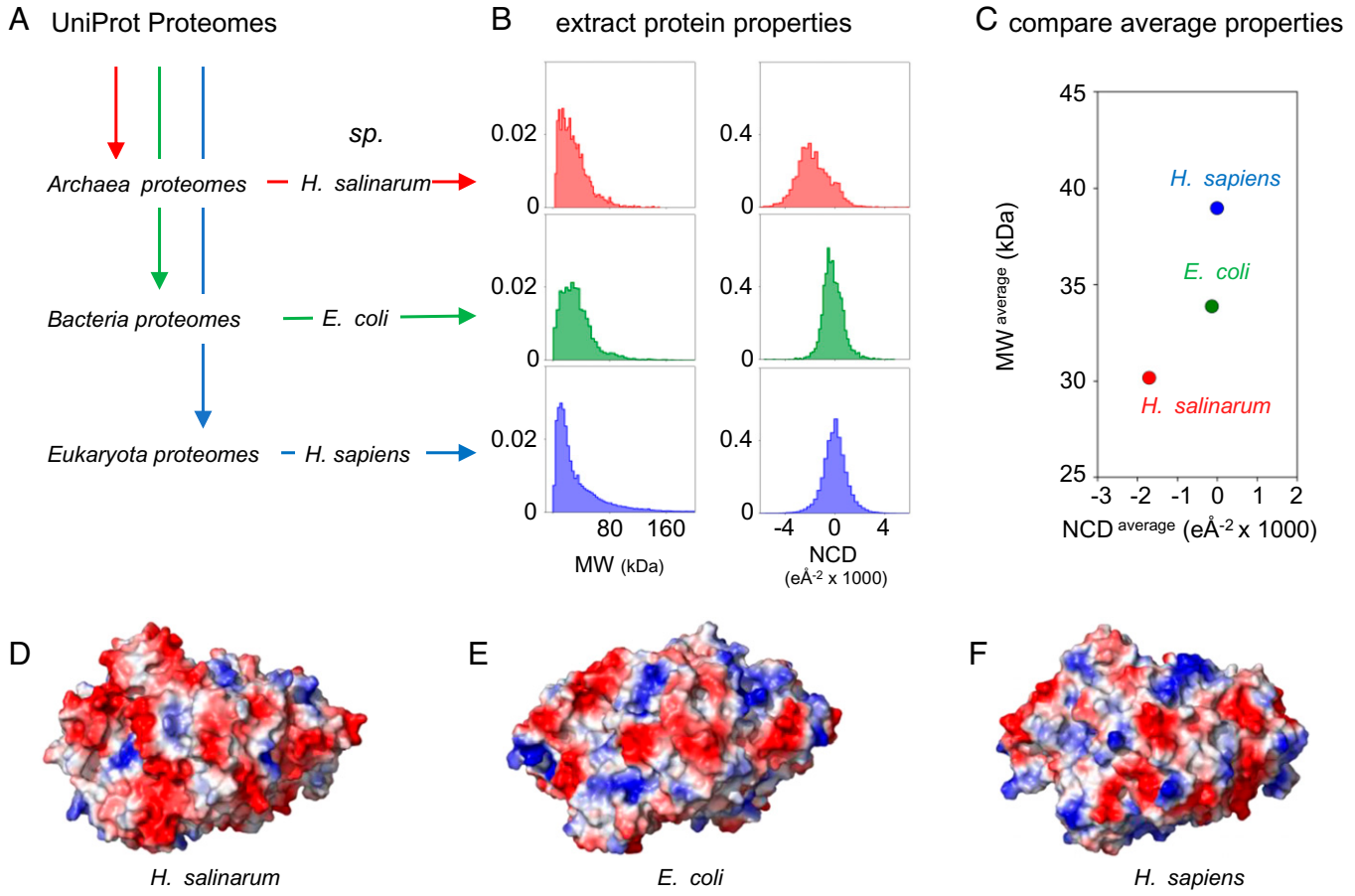

Fig. 1. Outline of approach. (*A*) The UniProt Proteomes of *Archaea*, *Bacteria*, and *Eukaryota* were sampled and the physicochemical properties of each of their proteins calculated (*Methods*). From this dataset, containing ~10,000 proteomes, any species can be further analyzed with respect to its detailed protein properties. (*B*) The distributions of protein MW and protein NCD for the proteomes of *H. salinarum*, *E. coli*, and *H. sapiens*. (*C*) Plot of the average MW and NCD values, derived from the protein distributions in *B*. This two-dimensional representation shows clear separation of the three species. Corresponding plots can be obtained for all UniProt Proteomes and, similarly, for all NCBI Assemblies (*Methods*; *SI Appendix*, Fig. S1; and Fig. 2). (*D*–*F*) Generally, the protein-surface charge of individual proteins follows closely that of the proteome average (*SI Appendix*, Fig. S19). Shown are surface charge potentials in vacuum (red for negative and blue for positive) of ribonucleotide reductase orthologs in *H. salinarum* (UniProt Q9HMU3 modeled on PDB 5im3), *E. coli* (PDB 2xap), and *H. sapiens* (PDB 3hnc).

various biotopes and functional specializations. Given that the data cover ~18,000 organisms in all kingdoms of life, we focus below on a few representative examples of divergent optimizations and refer the specialist readers to our proteome explorer website for more specific analysis (https://proteome-explorer.herokuapp.com/).

Results

Organisms in the Physicochemical Coordinate Space. It is well known that cellular function relies on a complex network of specific protein-binding events, the details of which can be mapped out from the conservation of protein surfaces (24–27). Here, we explore instead what can be learned from the ubiquitous protein–protein interactions that are bound to be in competition with specific binding. As outlined above, a principal determinant for the diffusive interactions with the cellular interior is the protein net-charge density (4). To illustrate our approach, we start by comparing the net-charge densities (NCD) of each protein in the evolutionary divergent model organisms *Halobacterium salinarum* (halophilic archaeon), *E. coli* (gram-negative bacterium), and *Homo sapiens* (mammal) (*Methods*). The results show that the net-charge densities adhere to relatively narrow distributions centered around negative mean values (Fig. 1). These narrow feature distributions, which conveniently apply also to the other physicochemical observables analyzed in this study, allow us to simplify the comparison of the three organisms by reducing their proteomes to a set of average coordinates (*SI Appendix*, Table S1 and Figs. S1–S3). As base coordinates, we chose first the average net-charge density ($NCD^{average}$) and average protein size (average molecular weight [$MW^{average}$]), since these features are found to influence the intracellular protein diffusion experimentally (4, 13). Plots of $NCD^{average}$ vs. $MW^{average}$ show clear separation between the three model organisms, indicating already at this level a physicochemical diversification of their proteomes (Fig. 1 and *SI Appendix*, Fig. S4). Most apparently, $NCD^{average}$ of *H. salinarum* has a value ~10 times more negative than observed for *E. coli* and *H. sapiens* (Fig. 1). This simple procedure can be extended to any other organism whose proteome is represented in UniProt (~10,000) or, with minor changes, in the National Center for Biotechnological Information (NCBI) Assemblies database (~18,000; *Methods* and *SI Appendix*, Fig. S1). Since the two datasets are found to yield very similar results (*SI Appendix*, Fig. S5), we focus our analysis on the latter, which offers the largest taxonomic coverage.

The picture becomes even more persuasive upon inclusion of all organisms, revealing the breadth and diversity of the physicochemical landscape (Fig. 2). *Archaea*, *Bacteria*, and *Eukaryota* not only occupy different physicochemical space, but also their

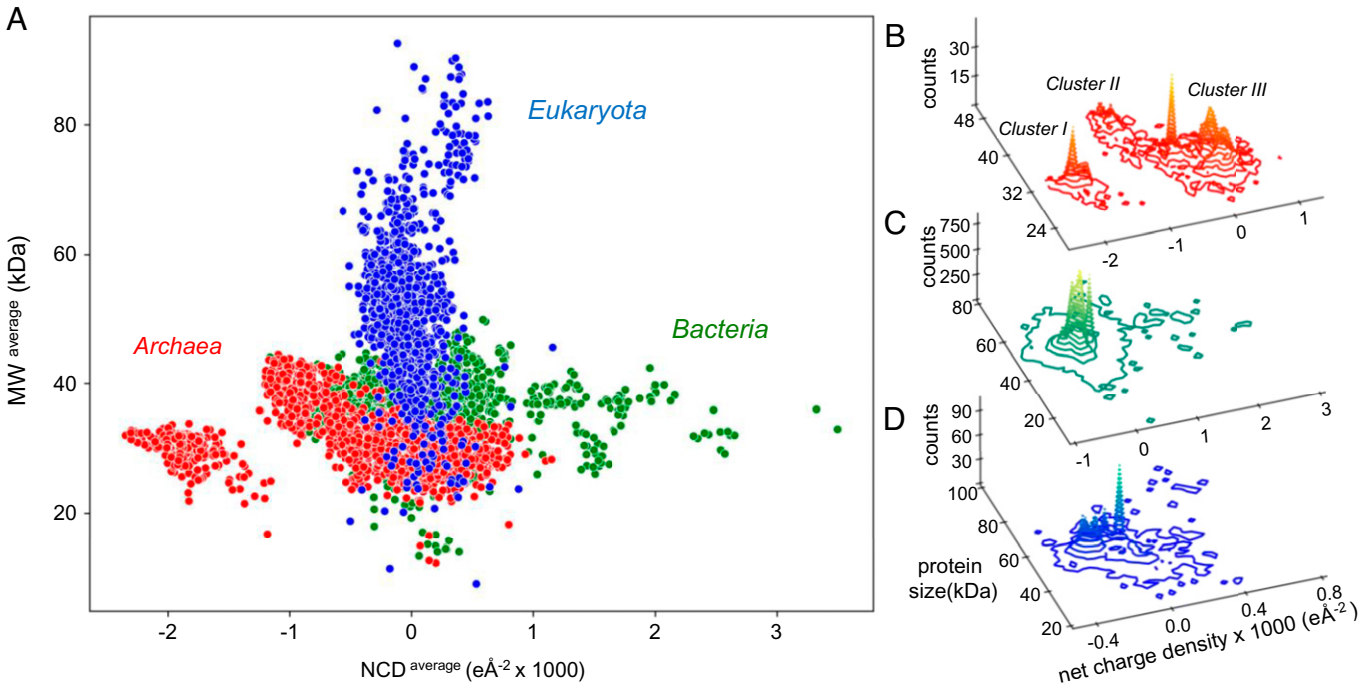

Fig. 2. The average proteome properties of the three kingdoms of life. (*A*) Plot of $NCD^{average}$ vs. $MW^{average}$ based on 2,223 archaeal (red), 13,652 bacterial (green), and 2,004 eukaryotic (blue) NCBI Assemblies. The data show clear divergence of proteome properties across the different organisms. (*B–D*) Density maps of the *Archaea*, *Bacteria*, and *Eukaryota* datasets, revealing detailed features in the physicochemical landscapes. As outlined in Figs. 3–6, the distinct peaks and clusters in these landscapes indicate various types of functional optimization at average protein level. For interactive versions of *B–D*, see Movies S1–S3.

distributions are markedly different in shape. While *Archaea* and *Bacteria* spread over a broad range of $NCD^{average}$ values, *Eukaryota* occupy a relatively narrow band (Fig. 2). Conversely, *Archaea* and *Bacteria* are constrained to a relatively narrow range of $MW^{average}$, whereas *Eukaryota* shows a much larger span that is also shifted to higher mass. The results underline that the properties of the variable protein surfaces are by no means random, but are biased to distinct regions of the parameter space and bear all the hallmarks of biological optimization.

Archaea and the Halophilic Niche. From the pattern of archaeal proteome properties, three subpopulations are immediately recognized (Fig. 2). The picture clarifies further upon coloring the data according to organism lineage, i.e., phylum, class, order, family, genus, and species (*Methods* and Fig. 3). Distinction of the cluster at high-negative $NCD^{average}$ is clear-cut (cluster 1), and its members represent two taxonomic classes of *Euryarchaeota*, i.e., *Halobacteria* and *Nanohaloarchaea*. Both taxa inhabit hypersaline environments (28), and their extremely negative proteomes have previously been linked to osmotic pressure adaptation (29). Essentially, these archaea prevent water loss to the hypersaline environment by accumulating multimolar cytosolic concentrations of electrostatically screening KCl and NaCl. This, in turn, requires that the protein surfaces become more negative to maintain viable repulsion and prevent full-scale aggregation (1, 30). The other two archaeal clusters are less clearly separated in the $NCD^{average}$–$MW^{average}$ projection but can be further deconvoluted by Gaussian mixture models (*SI Appendix*, Fig. S7) or principal component analysis (PCA) based on a larger set of physicochemical properties (*SI Appendix*, Fig. S8). Cluster 2 brings together an elusive group of marine archaea for which taxonomic information is yet scarce (31, 32). One-quarter of its members are currently listed under the class *Candidatus Poseidoniia* (former Marine Group II), while two-thirds of its members are associated with unclassified archaea, under NCBI:txid2026739 (Fig. 3). Another conspicuous feature of cluster 2 is that it displays $NCD^{average}$ values intermediate between the extreme halophiles and other organisms, indicating a separate line of functional adaptation (1). Finally, cluster 3 is centered around $NCD^{average} = 0$ and holds the main collection of archaeal taxa, including the recently described phyla *Thaumarcheota*, *Crenarchaeota*, and *Candidatus Lokiarchaeota* (33, 34) (Fig. 3). This orderly separation underlines that $NCD^{average}$ and $MW^{average}$ do not drift randomly during the course of evolution, but closely follow the organism lineage and habitat adaptation.

Bacteria and the Endosymbiotic Niche. A striking feature of the bacterial subset is the long tail of net-positive proteomes (Fig. 2). However, this anomalous positive charge is tied to the identity of these organisms and their lifestyle. The bacteria observed above $NCD^{average} = 0.001$ $eÅ^{-2}$ are all endosymbionts, such as *Buchnera*, *Candidatus Hodgkinia*, and *Blattabacterium*, which exclusively live inside insect cells (35–37). The mutualistic relationship imposes a purifying selection that has led to reduced proteomes with skewed charge distributions (38, 39). Consistently, these small proteomes are enriched in the positively charged proteins targeted to nucleic-acid binding in replication, transcription, and translation (*SI Appendix*, Fig. S9), while the "missing" negatively charged proteins are imported from the host (40). The other *Bacteria* exhibit a rather uniform scatter in the $NCD^{average}$–$MW^{average}$ projection, which is centered at a slightly negative value and lacks the separated subclusters observed for *Archaea* (Fig. 2). Despite the extensive overlap in this distribution, remarkable segregation still emerges upon linking the data to taxonomy. For the five largest phyla, i.e., *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, and *Tenericutes*, the data reveal five partly overlapping

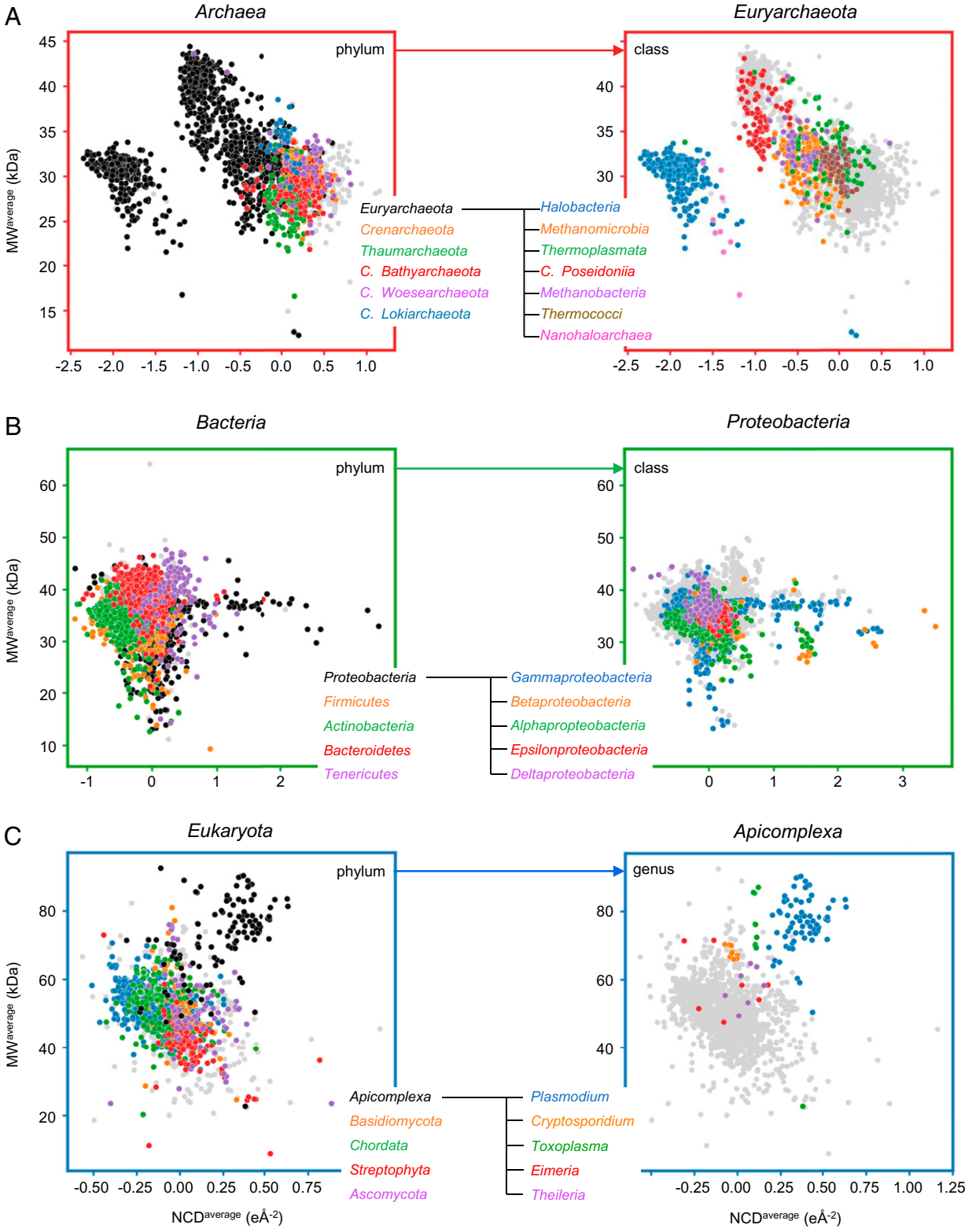

Fig. 3. $NCD^{average}$ vs. $MW^{average}$ maps of *Archaeal*, *Bacterial*, and *Eukaryotic* NCBI Assemblies where the individual organisms are colored according to taxonomic assignment. *A–C*, *Left* show the kingdom phyla, and *A–C*, *Right* show divisions into selected lower taxonomic levels, i.e., class or genus. Systematic biases appear at all taxonomic levels in terms of $NCD^{average}$ and $MW^{average}$, and these results can be reproduced in the smaller dataset based on UniProt Proteomes (*SI Appendix*, Fig. S6). (*A*) The five most abundant phyla of *Archaea*, where the selected phylum *Euryarchaeota* (black) contains members of all archaeal clusters (*SI Appendix*, Figs. S7 and S8). Upon division of *Euryarchaeota* into classes, it emerges that exclusively *Halobacteria* and *Nanohaloarchaea* make up cluster 1, *Candidatus Poseidoniia* maps to cluster 2, while all other classes of *Euryarchaeota* group in cluster 3. (*B*) The five most abundant phyla of *Bacteria*, where the largest phylum *Proteobacteria* (black) is further divided into classes. A notable feature is that the various classes of endosymbionts with markedly reduced number of genes are separated from other bacteria by having positive $NCD^{average}$ values (*SI Appendix*, Fig. S15). Conversely, assemblies of some *Mycoplasma* containing very small proteins are seen at $MW^{average} < 16$ kDa (*SI Appendix*, Fig. S18). (*C*) The five most abundant phyla of *Eukaryota*, all of which are biased to different regions of the property map. The most patent segregation involves *Apicomplexa* at high $NCD^{average}$ and $MW^{average}$ values. Upon dividing *Apicomplexa* into genus, this divergence of proteome properties is most pronounced for *Plasmodium*, containing the malaria parasites. Other examples of taxonomic segregation of eukaryotes in the $NCD^{average}$–$MW^{average}$ projection are shown in *SI Appendix*, Fig. S11.

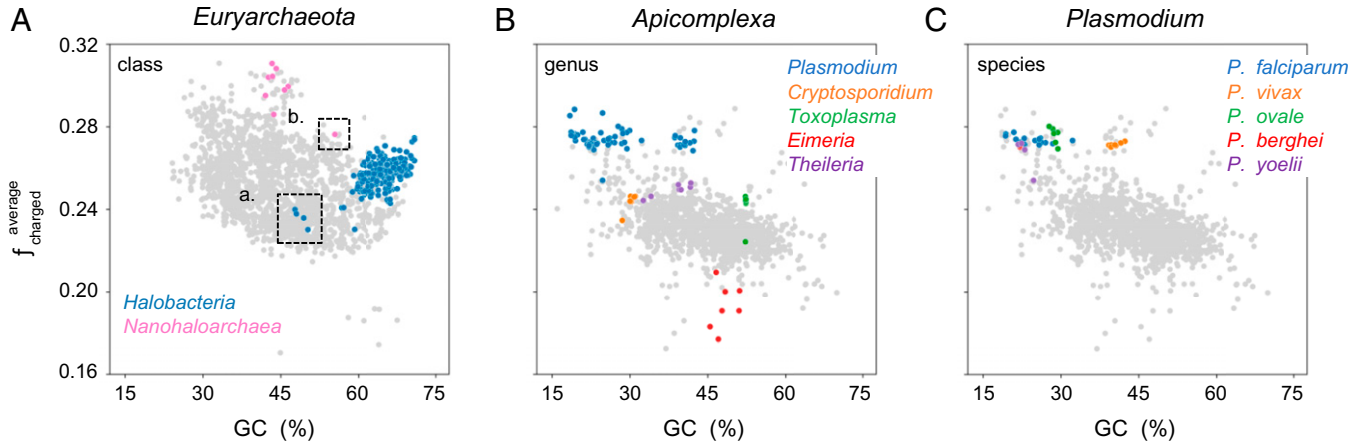

Fig. 4. Enhanced resolution by inclusion of the parameters GC and $f^{average}_{charged}$. (*A*) Separation of the classes *Halobacteria* and *Nanohaloarchaea* in cluster 1 of *Archaea* shows that adaptation to ionic-liquid environments is independent of GC content and prompts questions about convergent or divergent evolution. The *Halobacteria* outliers at intermediate GC (box a) point to the morphologically distinct representatives of *H. walsbyi* (47). The offset *Nanohaloarchaea* observation at 56% GC (box b) is the species *Candidatus nanosalinarum*, one of the first described members of the class and found in surface waters of a hypersaline Australian lake (28). (*B*) Genera within the phylum *Apicomplexa* occupy different regions of plane, including both extremes of the $f^{average}_{charged}$ dimension. (*C*) Remarkably, the genus *Plasmodium* can in turn be split into five of its species, which appear separated into narrow bands of GC content.

yet distinct clusters (Fig. 3). Each of these phyla further splits up into lower taxonomic levels with well-separated proteome properties that extend continuously down to bacterial classes (https://proteome-explorer.herokuapp.com/).

Eukaryotes and the Plasmodium Parasite. The overall proteome features of *Eukaryota* differ in two ways from those of *Archaea* and *Bacteria* (Fig. 2). Eukaryotes not only have adapted protein sizes that are 64% larger, but also show a cross-species $MW^{average}$ variation much greater than observed for the prokaryotes. At the extremes, we find the clusters of *Streptophyta* (plants) and *Apicomplexa* (parasitic protozoa) centered around $MW^{average} = 40$ and 80 kDa, respectively (Fig. 3). Contrasting the flexibility in size, *Eukaryota* are relatively restricted in terms of $NCD^{average}$: Their values are confined to a narrow interval between -10^{-3} and $+10^{-3}$ eÅ^{-2}, compared to a five-times larger spread for the prokaryotes (Fig. 2). Even so, *Eukaryota* reveal one marked division: A subset of species is offset to positive $NCD^{average}$ values and extreme $MW^{average}$ in Fig. 3 *C*, *Right*. Intriguingly, this minor cluster holds the phylum *Apicomplexa*, including the genus *Plasmodium* and the malaria parasites (41–43). The largest phyla in the main eukaryote cluster are *Basidiomycota* (filamentous fungi), *Chordata*, *Streptophyta* (plants), and *Ascomycota* (sac fungi) (Fig. 3), where the distinct peak in the $NCD^{average}$–$MW^{average}$ topology is caused by the sheer number of species in the latter (Fig. 2). *Eukaryota* maintain also orderly physicochemical separation at lower taxonomical levels, as observed for *Archaea* and *Bacteria* (*SI Appendix*, Fig. S10).

Additional Separation by Inclusion of Guanine–Cytosine Content and Fraction of Charged Residues. Although the distinction of proteome properties in Figs. 2 and 3 is indeed surprisingly detailed, it remains limited to two variables alone. The question is then, What happens if the analysis is extended to include additional properties of physiological interest? One such property is the fraction of charged residues ($f^{average}_{charged}$), and another is the enigmatic guanine–cytosine (GC) content (*SI Appendix*, Table S1) (44, 45). Although the latter is not linked to protein identity per se, it represents a key modulator of the physicochemical behavior of DNA that also varies across organisms in a phylogenetically consistent way (46). The effect of this parameter change is illustrated as follows: For *Archaea*, the $f^{average}_{charged}$– GC projection reveals additional complexity within the halophilic cluster 1, where *Halobacteria* cluster at the high end of the GC distribution, while *Nanohaloarchaea* occupy more moderate GC values (Fig. 4). However, there are some notable outliers at intermediate GC in both classes. Among *Halobacteria*, the outliers correspond to different assemblies of *Haloquadratum walsbyi* with peculiar square-shaped cells (47), and the single nanohaloarchaeon outlier represents one of the first described members of its class (28) (Fig. 4). The separation of the two classes implies either that 1) the highly negative proteomes of cluster 1 have arisen through convergent evolution or that 2) their GC content has diverged. Irrespectively, the organized partitioning in the plot emphasizes that the proteome properties do not drift randomly, but adjust to specific optimized values. Corresponding separations within *Bacteria* are shown in *SI Appendix*, Fig. S11, where, e.g., *Proteobacteria*, *Firmicutes*, and *Bacteroidetes* exhibit convergent proteome properties with divergent GC content. Within *Eukaryota*, the change of parameters reveals an orderly division of the phylum *Apicomplexa,* where the genera containing the most members split up into well-defined clusters. In turn, finer taxonomic differences are detected across several species of *Plasmodium*, which can be isolated due to their diverging GC contents (Fig. 4).

Separations Relating to Preferred Temperature and Morphological Features. It is well established that adaptations to extreme habitats and acquisition of certain cellular features often involve convergent evolution. One such adaptation is the ability to sustain extreme temperatures, where the bacterial proteomes display significant physicochemical preferences (Fig. 5 and Movies S4 and S5). Upon coloring the bacterial species according to their growth temperatures (T_{growth}), the thermophiles with $T_{growth} > 60$ °C cluster around $NCD^{average} = 0$, low GC, and high $f^{average}_{charged}$ (Fig. 5 and *SI Appendix*, Fig. S12). Conversely, the cold-adapted psychrophiles with $T_{growth} < 15$ °C show more negative $NCD^{average}$, high GC, and low $f^{average}_{charged}$ (Fig. 5 and *SI Appendix*, Fig. S12). A notable outlier is the model organism for thermophilic bacteria, *Thermus thermophilus* (48), which possesses relatively high GC and just moderate $f^{average}_{charged}$ (Fig. 5 and *SI Appendix*, Fig. S12). To exemplify proteome bias linked to morphological classification, we use the

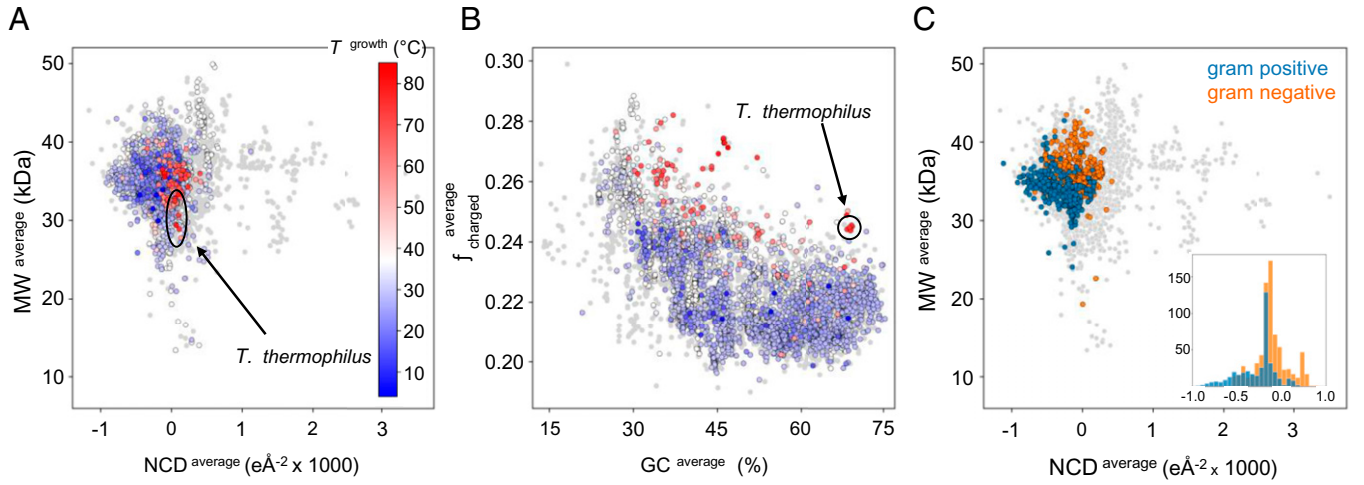

Fig. 5. Physicochemical differences related to organism lifestyle and morphology. (*A*) The $MW^{average}$–$NCD^{average}$ projection shows separation between thermophilic (T_{growth} > 60 °C) and mesophilic/psychrophilic (T_{growth} < 20 °C) bacteria, where the latter have more negatively charged proteomes. (*B*) At any level of GC, thermophiles show generally higher $f_{charged}^{average}$ values than the mesophiles and psychrophiles, consistent with previous notions. The model organism *T. thermophilus* (n = 6) stands out from the other thermophiles by having remarkably high GC (*SI Appendix*, Fig. S12). (*C*) The differences related to cell-wall exposure, where the gram-positive bacteria are generally offset to lower $MW^{average}$ and $NCD^{average}$.

"gram staining," by which bacteria are grouped according to their cell wall exposure (49). Separations occur here already in the $NCD^{average}$–$MW^{average}$ plane where the gram-positive bacteria show proteomes that are overall more negatively biased (Fig. 5). Clues to this difference are again found in their physiological adaptation: gram-positive bacteria rely generally on thick cell walls without external membranes, coupled to turgor pressures that are considerably higher than those of their gram-negative relatives (50, 51).

Uniform Manifold Approximation and Projection: Making the Most of the Data. Uniform manifold approximation and projection (UMAP) is a dimensionality reduction technique that has successfully been used to visualize data from single-cell mass cytometry and RNA-sequencing studies (52, 53). In this study, UMAP projections based on $NCD^{average}$, $MW^{average}$, GC, and $f_{charged}^{average}$ show that the global separations of *Archaea*, *Bacteria*, and *Eukaryota* are basically retained, but with improved congruency (Fig. 6). For *Archaea*, UMAP confirms the three archaeal clusters in Fig. 6 and better resolves the boundary between clusters 2 and 3 (Fig. 6). At lower taxonomic levels, the classes *Halobacteria* and *Candidatus Poseidoniia* within the phylum *Euryarchaeota* are moved farther apart, and even subtler differences between the orders of *Methanomicrobia* emerge (*SI Appendix*, Fig. S13). For *Bacteria* and *Eukaryota*, the UMAP projections largely improve the separation of the main phyla, reducing the overlap observed in Fig. 3 (*SI Appendix*, Figs. S14–S17).

The power of the UMAP projections, however, is best demonstrated by examining lower ranks in the taxonomic hierarchy. For example, the well-studied class *Gammaproteobacteria*, containing many medically, ecologically, and technologically important groups of bacteria (54, 55), reveals clear differences in physicochemical properties between closely related taxa down to genus level (Fig. 6). In particular, the pathogen-containing *Klebsiella* (56) stand out from their sibling genera *Escherichia*, *Enterobacter*, and *Salmonella*. These genera also provide examples of broad property spreads that possibly relate to their variable mosaic genomic structures, accounting sometimes for > 50% of the protein content (57, 58). Correspondingly distinct separations are observed for *Eukaryota*, where the lineages for pathogens and other biologically interesting organisms often run unbroken to individual species (*SI Appendix*, Figs. S16 and S17). Although UMAP embeddings offer remarkable resolving power, the physicochemical properties underlying the separations are obscured by the nonlinear nature of the method (53). Even so, the missing information can strictly be traced back to the input parameters $NCD^{average}$, $MW^{average}$, GC, and $f_{charged}^{average}$ (*Methods*), making UMAP a quick and sensitive tool for initial exploration of proteome differences across large datasets.

Implications from Outliers: Taxonomic Misassignments and Proteome-Quality Check. Because our analysis is based on average protein properties rather than on specific sequence details, the position of an organism in the physicochemical space is relatively robust and insensitive to variability in the individual genes (*SI Appendix*, Figs. S5 and S6). The question is then what the odd outliers that still emerge in the datasets represent. Inspection of the halobacterial assemblies that are offset to near-neutral $NCD^{average}$ values (Fig. 3) shows that these either contain large numbers of pseudogenes or lack support for conclusive taxonomic assignment (*SI Appendix*). Although it remains to be examined what further categories of outliers there are, the ease with which they can be spotted offers a quick quality control of gene assemblies and taxonomic assignment (*SI Appendix*, Fig. S18).

Discussion

The physicochemical features of biomolecules are key to cellular viability, and we demonstrate here that simplistic analysis of protein-surface properties alone can be used for taxonomical distinction of organisms across all kingdoms of life. In essence, the protein surfaces of the various organisms display relatively narrow net-charge density distributions, the average values of which ($NCD^{average}$) diverge in an evolutionarily consistent way (Fig. 1). The driving force behind this divergence is possibly the need to keep the proteome optimally fluid under different adaptive pressures (1, 4, 14–16). A telling example is the extremely negative $NCD^{average}$ values of *Halobacteria*, where the gain of extra repulsive charge opposes proteome aggregation in cells with high concentrations of screening salt (1, 19–21)

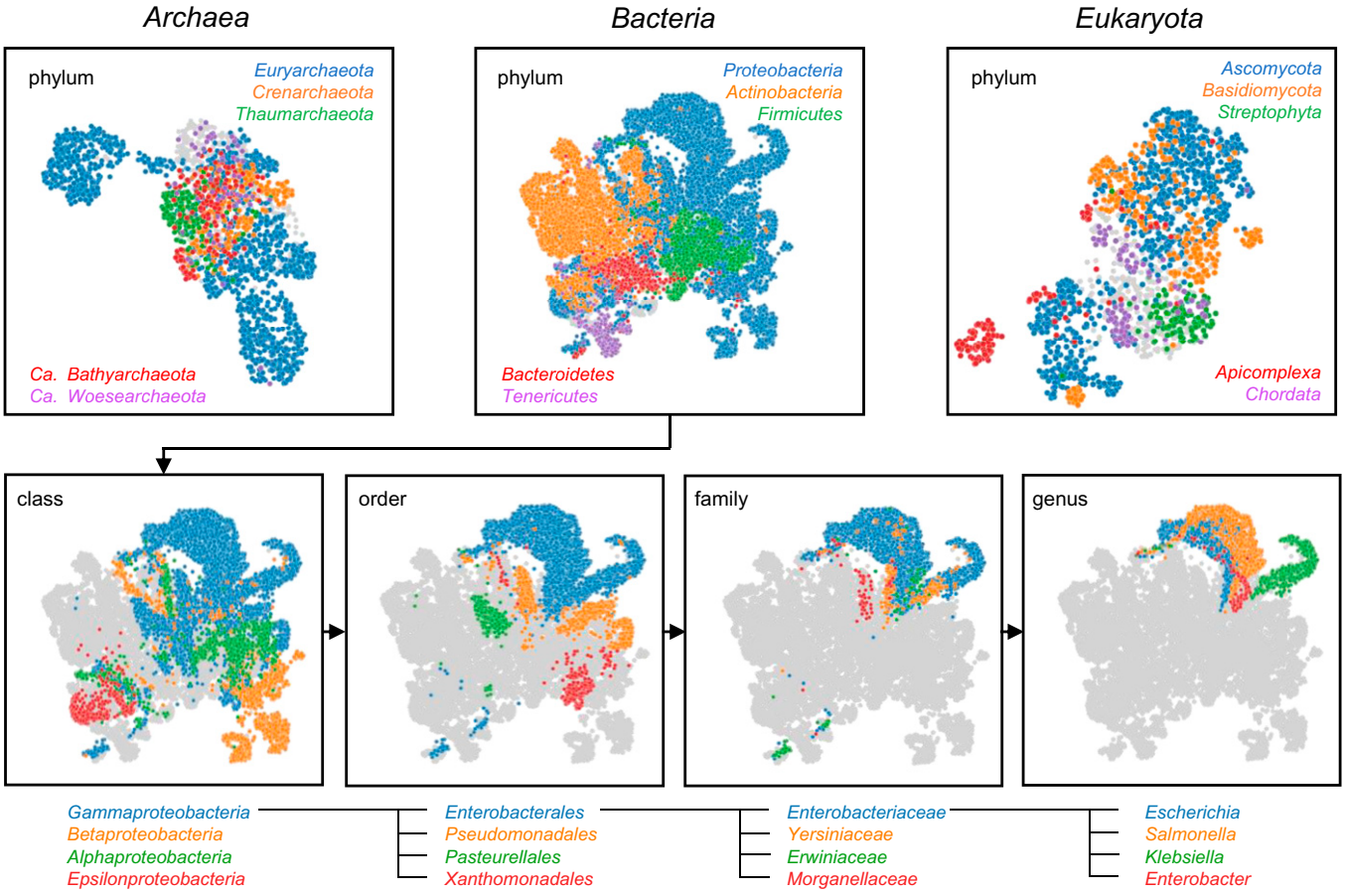

Fig. 6. UMAP representations of the three kingdoms of life, based on the physiochemical properties $NCD^{average}$, $MW^{average}$, GC, and $f^{average}_{charged}$. *Top* panels mark the five largest phyla (Fig. 3), where splitting into lower taxonomic levels is shown only for *Bacteria* (*Bottom*). The UMAP separations at genus level are largely, but not only, linked to differences in GC (*SI Appendix*, Fig. S14).

(Fig. 2). However, such orderly divergence of proteome $NCD^{average}$ is not limited to extremophiles, but is observed for most organisms in minimalist $MW^{average}$–$NCD^{average}$ plots (Fig. 3). Additional adaptive cues are revealed by extending the physicochemical space to include other variables of interest, such as the proteome $f^{average}_{charged}$ and the genomic GC (Fig. 4). Organism divergence is then resolved at most taxonomic levels, ranging from distinct clustering of phyla down to local separations of genus and even individual species (Fig. 3 and *SI Appendix*, Fig. S16). Corresponding patterns of separation emerge also on a collective scale upon grouping the prokaryotic organisms according to growth temperatures (T_{growth}), cell-wall characteristics, and endosymbiotic lifestyle, where the preference for a certain set of proteome features spans multiple taxa (Fig. 5 and *SI Appendix*, Fig. S15). This convergence across distantly related organisms suggests that the proteome properties are not coupled to the lineage or taxonomy per se, but dictated by optimization of cellular function to a given set of habitat conditions or morphological constraints. In other words, adaptation to a certain niche involves not only the acquisition of specific sets of functional loci, but also adjustments of the proteome properties as a whole, expanding the notion of "habitat genomes" (59–61). Since the present analysis targets mainly the variable protein surfaces, it provides evolutionary information orthogonal to that obtained from conventional sequence alignments (62, 63). Similar attempts to classify organisms by their protein features have previously been made on smaller subsets of organisms, using isoelectric focusing (23), mass spectrometry (64, 65), and computational estimates of protein isoelectric point (16, 66, 67). The results from these studies are overall consistent with the observations presented here, underlining that optimization of the hypervariable parts of the protein surfaces occurs generally during evolution.

Regarding the applications of protein-property mapping, it can readily aid 1) classification (Figs. 3–6 and *SI Appendix*, Figs. S10, S11, and S13), 2) quality control of deposited genomes, and 3) determination of variability and constraints in physicochemical adaptations (*SI Appendix*). With respect to the latter, one question concerns the very limits for life in extreme environments (68), with implications for astrobiology (69) and primordial evolution (70). A related issue is why there is such a diversity of alternative adaptations for a given biotope. One example is the cluster-2 *Archaea* with $NCD^{average}$ values intermediate between those of most other marine organisms and the brine-adapted *Halobacteria* (Fig. 3). Despite that, the cluster-2 *Archaea* are by no means extremophiles, but appear exclusively in samples from the surface regions of temperate oceans around the globe (31, 32). These archaea may then employ their own physiological strategy to survive in this competitive niche, perhaps by using an energy-conserving salt-in strategy similar to that of *Halobacteria* (71). If so, their internal salt concentrations will match that of the oceans at around 0.5 M NaCl, where the functional electrostatic interactions are suppressed due to excessive ion screening (1, 72, 73). Interestingly, this seemingly unfavorable cytosolic situation applies also to any prebiotic chemistry that has occurred in marine environments (74), emphasizing the need to better understand these alternative, and perhaps ancient, adaptive solutions. To this end, it

remains to see whether the deviating proteome properties of some pathogens and parasites can inform the optimization of therapeutic targeting. One example is the positive $NCD^{average}$ bias of the malaria parasites (*Plasmodium*), which clearly contrasts with that of the human host (Figs. 3, 4, and 6). Also, we note the SOD1 mutations associated with the neurodegenerative disease amyotrophic lateral sclerosis (ALS) preferentially reduce the protein's net-negative repulsive charge (6) and that electrostatic interactions are bound to play a central role in the formation of biological condensates (75, 76).

When it comes to protein compatibility across species, information can be sought in the physicochemical profiles of organisms involved in lateral gene transfer events. As one example, the suggested donor for transfer of ribonucleotide reductases to *Halobacteria* is the halophilic bacterium *Salinibacter ruber* (77). Not only do these organisms share the same habitat, but also the proteome $NCD^{average}$ of *S. ruber* is on the extreme negative side of the bacterial distribution and thus very close to *Halobacteria* (*SI Appendix*, Fig. S20). Such matching of proteome properties in lateral gene transfer is indeed expected to favor functionality, which can be a useful consideration in synthetic biology (78) and in cellular protein engineering in general. Along this line, we observe by in-cell NMR that transfer of the human HAH1 protein into *E. coli* causes a marked arrest of its rotational diffusion, which predictably restores when the protein is mutated to have average surface properties similar to that of the *E. coli* homolog (4). While this single-protein case is easy to understand, the evolution of the proteome-wide average poses a more intricate problem. To acquire the halophilic character of *H. salinarum*, for instance, a cluster-3 archaeon must on average add ∼20 negative charges to each protein surface (*SI Appendix*, Fig. S19). Since the diffusive encounters between intracellular proteins seem to be at the verge of specific interaction (79), such an evolutionary pathway is expected to involve numerous new binding constellations and loss of functional colocalization (80). Also, it will cause substantial increase of the protein–counterion ratios, an evolutionary transition across the minimum in charge-screening length (1), and require large-scale change of strategy for maintaining viable osmotic pressure (81, 82). The improbable picture that emerges is an evolutionary pathway that goes uphill; i.e., the organisms' fitness decreases as charges accumulate on their protein surfaces, until a breakthrough allows them to colonize a hypersaline environment. Although the changes of proteome properties between divergent taxa are generally smaller, and lateral gene transfer will speed up the transformation, the halophile example illustrates the extent of the problem and points to a question in cellular evolution that is not yet resolved. Regarding the time scales for physicochemical diversification, it is also of interest to establish the relative rates of $NCD^{average}$ and GC alterations. Candidates for such analysis are *Halobacteria* and *Nanohaloarchaea* with similar extreme $NCD^{average}$ values, but with GC of 65 and 45%, respectively (Fig. 4).

Obviously, the proteome-wide properties on which the present analysis focuses (Fig. 1) carry several simplifications and are by no means comprehensive. For example, we have left out the projections relating to surface hydrophobicity, despite that this parameter represents one of the basic forces controlling protein–protein interactions (4, 83, 84) and also indicates systematic cross-organism divergence complementary to that of charge (*SI Appendix*, Fig. S21 and Movie S6). The reason is that estimates of surface hydrophobicity from sequence data alone remain uncertain and are better saved for approaches based on high-resolution structures. Our analysis also does not cover to what extent the physicochemical flavor of the intracellular proteins responds to changes of expression patterns—e.g., in the cell cycle, various development phases, and stress responses—as well as the influence from the variable genomic composition within many bacterial genera (57, 58, 85). For instance, cytoplasmic acidification is reported to play a crucial role in mounting endogenous protein condensation and heat-shock response in yeast (86). Based on our current observations, the lowered pH is here expected to decrease the basal net-negative repulsion, favoring at the general level protein–protein association. To this end, the present study shows that the hypervariable protein surfaces exhibit formerly unappreciated signs of evolutionary control, possibly relating to their role in modulating diffusive protein encounters and intracellular solubility (1). While these signs may go unnoticed in conventional sequence alignment, they underline that global optimization of protein crosstalk is a universal determinant of cellular fitness. Thus, physicochemical profiling of proteins adds a perspective on cellular function, going beyond specific genes to global proteome features, and raises the question of whether there are even more layers in the molecular optimization of organisms waiting to be uncovered.

Methods

Datasets: NCBI Assemblies and UniProt Proteomes. Our knowledge of the proteins expressed by any given organism is constantly expanding, in pace with the growing number of sequenced genomes. To determine and cross-validate proteome properties across species, two complementary datasets were constructed over the course of 2 y: One was obtained from the NCBI in December 2019 and the other from UniProt Proteomes in October 2021. Both datasets were designed to meet certain quality criteria. From NCBI, we analyzed the GenBank annotations of all assemblies and excluded those flagged as partial, anomalous, or derived from large multi-isolate projects. Also, since bacterial sequences are vastly more abundant than those for archaea and eukaryotes, we limited our bacterial subset to complete genomes only. From UniProt, we took only reference proteomes into consideration, as this dataset also serves as quality control. Finally, records containing fewer than 50 proteins were disregarded across the board, limiting the number of proteomes in the NCBI and UniProt datasets to ∼18,000 and 10,000, respectively (*SI Appendix*, Fig. S1). The taxonomic lineages were then obtained for every proteome in these datasets through their NCBI TaxIds with TaxonKit, a command-line interface tool for handling NCBI taxonomy data (87), and these lineages were considered at six hierarchical levels, i.e., phylum, class, order, family, genus, and species.

Considering that the 10,000 UniProt proteomes are constructed from selected NCBI assemblies, it is notable that only 3,276 records are shared by the two datasets. There are at least two reasons for this partial overlap. First, as sequence data are in continuous development, some current UniProt Proteomes map to more recent NCBI assemblies that did not exist in 2019. Second, many bacterial UniProt proteomes are constructed from incomplete NCBI genomes at lower assembly levels, i.e., contigs and scaffolds, not present in our NCBI dataset, which included bacterial complete genomes only. A detailed comparison of how our two datasets were derived is presented in *SI Appendix*, Figs. S5 and S6.

Physicochemical Features, Residue Compositions, and Organism Profiles. The physicochemical properties of proteins were derived directly from their sequences, yielding a total of 11 features (*SI Appendix*, Table S1). Protein solvent-accessible surface area (SASA) was obtained from MW, using Miller's empirical equation (88). Protein net charge ($charge_{net}$) and the fraction of charged residues ($f_{charged}$) were derived from the number of acidic (D and E) and basic (R and K) residues in the individual protein sequences divided by sequence length ($l_{sequence}$). While $charge_{net}$ has relevance for the description of protein electrostatics (1), $f_{charged}$ has been reported as a controlling aspect in the context of protein hydration (89). SASA and $charge_{net}$ estimates were then combined to obtain protein NCD, which is a central parameter in the treatment of colloidal systems (1). Based on the general architecture of protein structures, this

derivation assumes that all charged residues are located outside the hydrophobic core and contained at the protein surface. As an estimate of protein hydrophobicity, we used the sequence fraction of the four most hydrophobic residues F, L, I, and V ($f_{\text{hydrophobic}}$). Finally, proteome size was defined by the number of annotated genes (n_{genes}). To obtain a corresponding physicochemical parameter for the DNA, we resorted to the whole-genome content of guanidine and cytosine bases (GC), as listed in the NCBI Genome Reports for prokaryotes and eukaryotes. Thus, our physicochemical properties define an 11-dimensional space, in which the various organisms are compared. In most figures, we base this comparison on the average features of all proteins in a given proteome, denoted with the superscript *average* (Figs. 1–5 and *SI Appendix*, Figs. S4–S7, S10–S15 and S17–S21).

As complementary analysis, organism coordinates were also obtained for residue composition, expressed as the proteome averages of the fractional abundances of the 20 naturally occurring amino acids. In other words, for each residue *i* we count the number of occurrences in a sequence and divide this by l_{sequence} to obtain 20 fractions, f_A, f_C … f_i such that $\sum_i^{20} f_i = 1$.

Notably, the proteomes selected for downstream analysis were considered in full. That is, they contain proteins of all sizes, including large proteins for which the Miller approximation of SASA performs worse (1), and subcellular localizations outside the cytosol, including membrane proteins. The skewing introduced by this simplification, however, remains small: Previous controls show that the physicochemical features of whole proteomes are essentially indistinguishable from those derived from cytosolic proteins of high abundance only (1). Except for GC, which is identical for both the NCBI and UniProt datasets (*SI Appendix*, Fig. S1), the exact score of each organism on every other feature can be different, as the two databases of origin rely on different methods and criteria to build their records. This allows us to evaluate the robustness of our results by comparing to what extent they can be reproduced with the two datasets (*SI Appendix*).

Gram Staining and Optimal-Growth Temperature. To see how features other than the organism's position in the taxonomic hierarchy are reflected in the physicochemical properties of bacteria, data relating to growth temperature (T_{growth}) and gram staining were obtained from BacDive (90). These records were merged with our dataset of bacterial physicochemical features, through NCBI TaxId, and yielded subsets of 1,146 entries with annotated gram stain and 7,780 entries with annotated growth temperatures.

Statistical Methods. Data analysis was carried out using custom Python code (version 3.6) and the scikit-learn module (91). Decomposition of the archaeal dataset into three clusters in the $\text{NCD}^{\text{average}}$–$\text{MW}^{\text{average}}$ projection was based on a Gaussian mixture model (GMM) with three subpopulations and tied covariance (91). PCA of the same dataset was performed on 9 of the 11 physicochemical features in *SI Appendix*, Table S1: l_{sequence} and SASA were excluded because of their inherent correlation with MW. PCA was also performed on all 20 features of amino acid composition. Finally, UMAP was performed as implemented in the Python module created by McInnes et al. (53). All reported results were based on the four features $\text{NCD}^{\text{average}}$, $\text{MW}^{\text{average}}$, GC, and $f_{\text{charged}}^{\text{average}}$. After exploration of several combinations of parameters, we opted for visualizations obtained with 100 neighbors and 0.99 minimum distance (53), where the "number of neighbors" affects how well the global structure is preserved, and the "minimum distance" affects the local overlap between observations in the resulting projection. PCA and UMAP were run on the archaeal, bacterial, and eukaryotic subsets separately, to avoid that the bacteria overshadow the other two kingdoms due to their sheer number.

Ribonucleotide Reductase Structures. As base for the comparison, we used the ribonucleotide reductase (RNR) structures for *H. sapiens* (Protein Data Bank [PDB] code 3hnc) and *E. coli* (PDB code 2xap). The corresponding RNR structures for *H. salinarum* and *S. ruber* were modeled on the homologous proteins from *Pseudomonas aeruginosa* (PDB code 5im3) and *Thermotoga maritima* (PDB code 1xjk), respectively. Both models were built on SWISS-MODEL (92) with default parameters, where the templates with highest sequence identity were selected.

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

ACKNOWLEDGMENTS. We thank Håkan Wennerström, Jens Danielsson, Peter Brzezinski, and Vicky Shingler for stimulating discussions. Funding was from the Knut and Alice Wallenberg Foundation (2017-0041) and the Swedish Research Council (2017-01517).

1. H. Wennerström, E. Vallina Estrada, J. Danielsson, M. Oliveberg, Colloidal stability of the living cell. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10113–10121 (2020).
2. R. D. Cohen, G. J. Pielak, A cell is more than the sum of its (dilute) parts: A brief history of quinary structure. *Protein Sci.* **26**, 403–413 (2017).
3. M. Gruebele, G. J. Pielak, Dynamical spectroscopy and microscopy of proteins in cells. *Curr. Opin. Struct. Biol.* **70**, 1–7 (2021).
4. X. Mu *et al.*, Physicochemical code for quinary protein interactions in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4556–E4563 (2017).
5. A. J. Guseman, S. L. Speer, G. M. Perez Goncalves, G. J. Pielak, Surface charge modulates protein-protein interactions in physiologically relevant environments. *Biochemistry* **57**, 1681–1684 (2018).
6. E. Sandelin, A. Nordlund, P. M. Andersen, S. S. Marklund, M. Oliveberg, Amyotrophic lateral sclerosis-associated copper/zinc superoxide dismutase mutations preferentially reduce the repulsive charge of the proteins. *J. Biol. Chem.* **282**, 21230–21236 (2007).
7. P. E. Schavemaker, W. M. Śmigiel, B. Poolman, Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. *eLife* **6**, 6 (2017).
8. G. Schreiber, A. R. Fersht, Energetics of protein-protein interactions: Analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478–486 (1995).
9. M. Kurnik, L. Hedberg, J. Danielsson, M. Oliveberg, Folding without charges. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5705–5710 (2012).
10. E. H. McConkey, Molecular evolution, intracellular organization, and the quinary structure of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3236–3240 (1982).
11. C. Eymann *et al.*, A comprehensive proteome map of growing Bacillus subtilis cells. *Proteomics* **4**, 2849–2876 (2004).
12. C. Pál, B. Papp, M. J. Lercher, An integrated view of protein evolution. *Nat. Rev. Genet.* **7**, 337–348 (2006).
13. S. Leeb, F. Yang, M. Oliveberg, J. Danielsson, Connecting longitudinal and transverse relaxation rates in live-cell NMR. *J. Phys. Chem. B* **124**, 10698–10707 (2020).
14. H. Langen *et al.*, Two-dimensional map of the proteome of Haemophilus influenzae. *Electrophoresis* **21**, 411–429 (2000).
15. L. P. Kozlowski, Proteome-pI: Proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116 (2017).
16. J. Kiraga *et al.*, The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8**, 163 (2007).
17. G. F. Weiller, G. Caraux, N. Sylvester, The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics* **4**, 943–949 (2004).
18. I. M. Karadzic, J. A. Maupin-Furlow, Improvement of two-dimensional gel electrophoresis proteome maps of the haloarchaeon Haloferax volcanii. *Proteomics* **5**, 354–359 (2005).
19. S. Paul, S. K. Bag, S. Das, E. T. Harvill, C. Dutta, Molecular signature of hypersaline adaptation: Insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* **9**, R70 (2008).
20. S. P. Kennedy, W. V. Ng, S. L. Salzberg, L. Hood, S. DasSarma, Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.* **11**, 1641–1650 (2001).
21. D. Madern, C. Ebel, G. Zaccai, Halophilic adaptation of enzymes. *Extremophiles* **4**, 91–98 (2000).
22. S. DasSarma, P. DasSarma, Halophiles and their enzymes: Negativity put to good use. *Curr. Opin. Microbiol.* **25**, 120–126 (2015).
23. A. Tebbe *et al.*, Life-style changes of a halophilic archaeon analyzed by quantitative proteomics. *Proteomics* **9**, 3843–3855 (2009).
24. I. M. Nooren, J. M. Thornton, Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991–1018 (2003).
25. E. M. Marcotte *et al.*, Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
26. I. Albert, R. Albert, Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* **20**, 3346–3352 (2004).
27. J. I. Garzón *et al.*, A computational interactome and functional annotation for the human proteome. *eLife* **5**, 5 (2016).
28. P. Narasingarao *et al.*, De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* **6**, 81–93 (2012).
29. R. Ghai *et al.*, New abundant microbial groups in aquatic hypersaline environments. *Sci. Rep.* **1**, 135 (2011).
30. A. Oren, Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes. *Front. Microbiol.* **4**, 315 (2013).
31. B. J. Tully, Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat. Commun.* **10**, 271 (2019).
32. E. F. DeLong, Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5685–5689 (1992).
33. K. Zaremba-Niedzwiedzka *et al.*, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
34. C. Brochier-Armanet, B. Boussau, S. Gribaldo, P. Forterre, Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* **6**, 245–252 (2008).
35. R. Gil, B. Sabater-Muñoz, A. Latorre, F. J. Silva, A. Moya, Extreme genome reduction in Buchnera spp.: Toward the minimal genome needed for symbiotic life. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4454–4458 (2002).
36. M. A. Campbell *et al.*, Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont Hodgkinia. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10192–10199 (2015).
37. Z. L. Sabree, S. Kambhampati, N. A. Moran, Nitrogen recycling and nutritional provisioning by Blattabacterium, the cockroach endosymbiont. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19521–19526 (2009).
38. G. M. Bennett, J. P. McCutcheon, B. R. MacDonald, D. Romanovicz, N. A. Moran, Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. *MBio* **5**, e01697-e14 (2014).

39. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
40. M. Balsera, J. Soll, B. Bölter, Protein import machineries in endosymbiotic organelles. *Cell. Mol. Life Sci.* **66**, 1903–1923 (2009).
41. S. M. Rich *et al.*, The origin of malignant malaria. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 14902–14907 (2009).
42. J. K. Baird, Neglect of Plasmodium vivax malaria. *Trends Parasitol.* **23**, 533–539 (2007).
43. W. E. Collins, G. M. Jeffery, Plasmodium malariae: Parasite and disease. *Clin. Microbiol. Rev.* **20**, 579–592 (2007).
44. F. Hildebrand, A. Meyer, A. Eyre-Walker, Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* **6**, e1001107 (2010).
45. A. Schmidt *et al.*, GC content-independent amino acid patterns in bacteria and archaea. *J. Basic Microbiol.* **52**, 195–205 (2012).
46. E. Haywood-Farmer, S. P. Otto, The evolution of genomic base composition in bacteria. *Evolution* **57**, 1783–1792 (2003).
47. H. Bolhuis *et al.*, The genome of the square archaeon *Haloquadratum walsbyi*: Life at the limits of water activity. *BMC Genomics* **7**, 169 (2006).
48. F. Cava, A. Hidalgo, J. Berenguer, Thermus thermophilus as biological model. *Extremophiles* **13**, 213–231 (2009).
49. M. R. J. Salton, The relationship between the nature of the cell wall and the Gram stain. *J. Gen. Microbiol.* **30**, 223–235 (1963).
50. G. Misra, E. R. Rojas, A. Gopinathan, K. C. Huang, Mechanical consequences of cell-wall turnover in the elongation of a Gram-positive bacterium. *Biophys. J.* **104**, 2342–2352 (2013).
51. A. M. Whatmore, R. H. Reed, Determination of turgor pressure in Bacillus subtilis: A possible role for K+ in turgor regulation. *J. Gen. Microbiol.* **136**, 2521–2526 (1990).
52. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
53. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [Preprint] (2018). https://doi.org/10.48550/arXiv.1802.03426 (Accessed 1 December 2021).
54. G. Rizzatti, L. R. Lopetuso, G. Gibiino, C. Binda, A. Gasbarrini, Proteobacteria: A common factor in human diseases. *BioMed Res. Int.* **2017**, 9351507 (2017).
55. H. H. Hau, J. A. Gralnick, Ecology and biotechnology of the genus Shewanella. *Annu. Rev. Microbiol.* **61**, 237–258 (2007).
56. E. Bouza, E. Cercenado, Klebsiella and enterobacter: Antibiotic resistance and treatment implications. *Semin. Respir. Infect.* **17**, 215–230.
57. E. N. Gordienko, M. D. Kazanov, M. S. Gelfand, Evolution of pan-genomes of *Escherichia coli*, *Shigella spp.*, and *Salmonella enterica*. *J. Bacteriol.* **195**, 2786–2792 (2013).
58. R. A. Welch *et al.*, Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020–17024 (2002).
59. E. F. Mongodin *et al.*, The genome of Salinibacter ruber: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18147–18152 (2005).
60. F. Dini-Andreote, F. D. Andreote, W. L. Araújo, J. T. Trevors, J. D. van Elsas, Bacterial genomes: Habitat specificity and uncharted organisms. *Microb. Ecol.* **64**, 1–7 (2012).
61. S. Sunagawa *et al.*; Tara Oceans coordinators, Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
62. G. C. Conant, P. F. Stadler, Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.* **26**, 1155–1161 (2009).
63. R. Sasidharan, C. Chothia, The selection of acceptable protein mutations. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 10080–10085 (2007).
64. S. Sauer, M. Kliem, Mass spectrometry tools for the classification and identification of bacteria. *Nat. Rev. Microbiol.* **8**, 74–82 (2010).
65. J. P. Dworzanski *et al.*, Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. *J. Proteome Res.* **5**, 76–87 (2006).
66. C. G. Knight, R. Kassen, H. Hebestreit, P. B. Rainey, Global analysis of predicted proteomes: Functional adaptation of physical properties. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8390–8395 (2004).
67. S. Nandi, N. Mehra, A. M. Lynn, A. Bhattacharya, Comparison of theoretical proteomes: Identification of COGs with conserved and variable pI within the multimodal pI distribution. *BMC Genomics* **6**, 116 (2005).
68. C. J. D. Lee *et al.*, NaCl-saturated brines are thermodynamically moderate, rather than extreme, microbial habitats. *FEMS Microbiol. Rev.* **42**, 672–693 (2018).
69. E. V. Pikuta, R. B. Hoover, J. Tang, Microbial extremophiles at the limits of life. *Crit. Rev. Microbiol.* **33**, 183–209 (2007).
70. C. R. Woese, On the evolution of cells. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8742–8747 (2002).
71. A. Oren, *Halophilic Microorganisms and Their Environments* (Springer Science & Business Media, 2006).
72. A. M. Smith, A. A. Lee, S. Perkin, The electrostatic screening length in concentrated electrolytes increases with concentration. *J. Phys. Chem. Lett.* **7**, 2157–2163 (2016).
73. A. A. Lee, C. S. Perez-Martinez, A. M. Smith, S. Perkin, Scaling analysis of the screening length in concentrated electrolytes. *Phys. Rev. Lett.* **119**, 026002 (2017).
74. W. Martin, J. Baross, D. Kelley, M. J. Russell, Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008).
75. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
76. J. Wang *et al.*, A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
77. D. Lundin, S. Gribaldo, E. Torrents, B.-M. Sjöberg, A. M. Poole, Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. *BMC Evol. Biol.* **10**, 383 (2010).
78. S. A. Benner, A. M. Sismour, Synthetic biology. *Nat. Rev. Genet.* **6**, 533–543 (2005).
79. T. Sikosek, H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface* **11**, 20140419 (2014).
80. E. A. Abbondanzieri, A. S. Meyer, More than just a phase: The search for membraneless organelles in the bacterial cytoplasm. *Curr. Genet.* **65**, 691–694 (2019).
81. J. van den Berg, A. J. Boersma, B. Poolman, Microorganisms maintain crowding homeostasis. *Nat. Rev. Microbiol.* **15**, 309–318 (2017).
82. M. C. Konopka, I. A. Shkel, S. Cayley, M. T. Record, J. C. Weisshaar, Crowding and confinement effects on protein diffusion in vivo. *J. Bacteriol.* **188**, 6115–6123 (2006).
83. L. A. Abriata, E. Spiga, M. D. Peraro, Molecular effects of concentrated solutes on protein hydration, dynamics, and electrostatics. *Biophys. J.* **111**, 743–755 (2016).
84. Q. Wang, A. Zhuravleva, L. M. Gierasch, Exploring weak, transient protein–protein interactions in crowded in vivo environments by in-cell nuclear magnetic resonance spectroscopy. *Biochemistry* **50**, 9225–9236 (2011).
85. H. Tettelin *et al.*, Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13950–13955 (2005).
86. C. G. Triandafillou, C. D. Katanski, A. R. Dinner, D. A. Drummond, Transient intracellular acidification regulates the core transcriptional heat shock response. *eLife* **9**, e54880 (2020).
87. W. Shen, J. Xiong, TaxonKit: A cross-platform and efficient NCBI taxonomy toolkit. bioRxiv [Preprint] (2019). https://doi.org/10.1101/513523 (Accessed 1 December 2021).
88. S. Miller, J. Janin, A. M. Lesk, C. Chothia, Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).
89. G. Ortega, T. Diercks, O. Millet, Halophilic protein adaptation results from synergistic residue-ion interactions in the folded and unfolded states. *Chem. Biol.* **22**, 1597–1607 (2015).
90. L. C. Reimer *et al.*, BacDive in 2019: Bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic Acids Res.* **47**, D631–D636 (2019).
91. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
92. A. Waterhouse *et al.*, SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).