Article

The proteome landscape of the kingdoms of life

https://doi.org/10.1038/s41586-020-2402-x

Received: 2 August 2019

Accepted: 27 April 2020

Published online: 17 June 2020

Check for updates

Johannes B. Müller[1,7], Philipp E. Geyer[1,2,7], Ana R. Colaço[3], Peter V. Treit[1], Maximilian T. Strauss[1,2], Mario Oroshi[1], Sophia Doll[1,2], Sebastian Virreira Winter[1,2], Jakob M. Bader[1], Niklas Köhler[4], Fabian Theis[4,5], Alberto Santos[3,6] & Matthias Mann[1,3]✉

Proteins carry out the vast majority of functions in all biological domains, but for technological reasons their large-scale investigation has lagged behind the study of genomes. Since the first essentially complete eukaryotic proteome was reported[1], advances in mass-spectrometry-based proteomics[2] have enabled increasingly comprehensive identification and quantification of the human proteome[3–6]. However, there have been few comparisons across species[7,8], in stark contrast with genomics initiatives[9]. Here we use an advanced proteomics workflow—in which the peptide separation step is performed by a microstructured and extremely reproducible chromatographic system—for the in-depth study of 100 taxonomically diverse organisms. With two million peptide and 340,000 stringent protein identifications obtained in a standardized manner, we double the number of proteins with solid experimental evidence known to the scientific community. The data also provide a large-scale case study for sequence-based machine learning, as we demonstrate by experimentally confirming the predicted properties of peptides from *Bacteroides uniformis*. Our results offer a comparative view of the functional organization of organisms across the entire evolutionary range. A remarkably high fraction of the total proteome mass in all kingdoms is dedicated to protein homeostasis and folding, highlighting the biological challenge of maintaining protein structure in all branches of life. Likewise, a universally high fraction is involved in supplying energy resources, although these pathways range from photosynthesis through iron sulfur metabolism to carbohydrate metabolism. Generally, however, proteins and proteomes are remarkably diverse between organisms, and they can readily be explored and functionally compared at www.proteomesoflife.org.

To collect a diverse set of representative organisms across the tree of life, we considered the availability of assembled genome sequences and the accessibility of cultured or tissue material, and included common model organisms for comparison. This resulted in 19 archaea, 49 bacteria and 32 eukaryotes—a total of 100 different species (Fig. 1a, b). We also added 14 viruses (Supplementary Table 1).

To obtain the proteomes of these extremely different biomaterials, we tested a number of extraction protocols and found that the in-StageTip (iST) protocol[10] was most universally applicable and allowed automated and highly reproducible sample preparation. We incorporated the latest advances into our workflow for high-resolution bottom-up proteomics, and implemented a recently developed chip-based method[11] (Fig. 1c–e). C_{18}-covered beads are replaced by a uniformly ordered and statically fixed micrometre-sized pillar structure[12] (Fig. 1d), leading to 2.5-fold improvements in coefficients of variation for peptide retention times and high interlaboratory reproducibility (Extended Data Figs. 1, 2a). For all prokaryotes we performed single-run mass spectrometry (MS) analyses, whereas we used a loss-less prefractionator[13] for the more complex eukaryotic samples.

We reasoned that our chip-based chromatographic method, combined with the very large data set of more than two million unique peptides, should be well suited to deep learning algorithms, which have recently been shown to be applicable to MS-based proteomics[14–16] (Extended Data Fig. 3). We developed a long short-term memory (LSTM) deep learning model with an interpretable attention layer to precisely predict chromatographic retention times, achieving a Pearson correlation of 0.990 (Extended Data Figs. 2b, 4). To test the model on a completely unknown proteome, we instructed the mass spectrometer to sequence peptides from *B. uniformis*, *Bacillus megaterium* or *Enterobacter aerogenes* only if they eluted in a narrow band around the retention times predicted by deep learning. This resulted in only slightly diminished proteome depths (at least 88% on the protein level), showing that these peptide properties were successfully modelled in silico (Fig. 2).

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]OmicEra Diagnostics GmbH, Planegg, Germany. [3]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany. [5]Technical University of Munich, Department of Mathematics, Garching, Germany. [6]Li-Ka Shing Big Data Institute, University of Oxford, Oxford, UK. [7]These authors contributed equally: Johannes B. Müller, Philipp E. Geyer. ✉e-mail: mmann@biochem.mpg.de

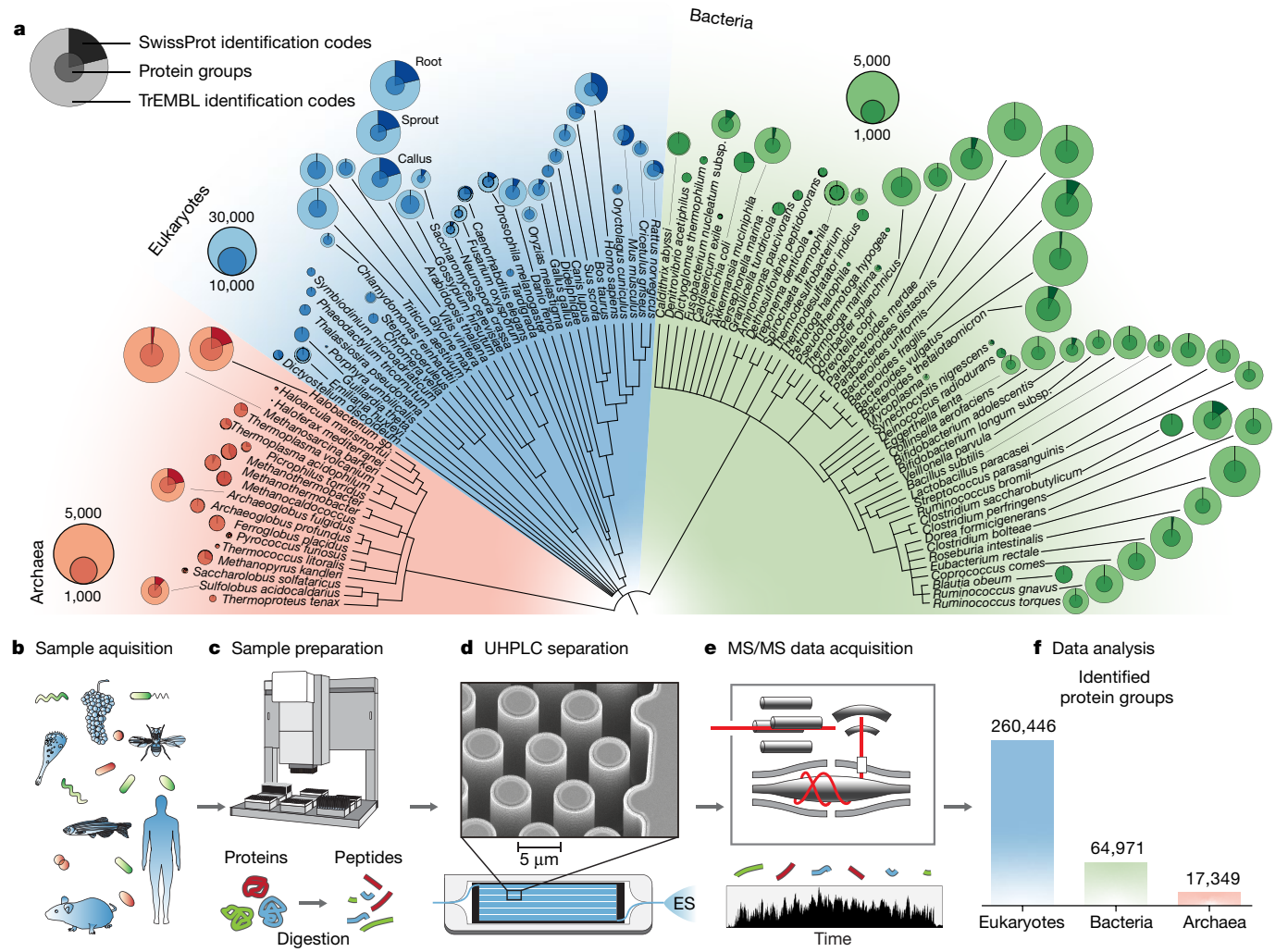


Fig. 1 | Collection of organism samples across the tree of life, and integration of the proteomic workflow. a, All organisms used herein were ordered and ranked on the basis of National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) taxonomy. Pie charts refer to the numbers of protein groups (proteins distinguishable by their identified peptides) and to database protein entries found here. **b**, **c**, The acquired samples were subjected to protein extraction and digestion into peptides for sample preparation. **d**, Peptides were separated using a silica-chip-based micropillar array column (µPAC) with etched pillar structures that are coated with C_{18}. UHPLC, ultra-high performance liquid chromatography. The magnification shows a scanning electron microscopy image of the pillar structures (adapted with permission from PharmaFluidics). **e**, Peptides were ionized by electrospray (ES) and analysed in a high-resolution mass spectrometer. **f**, Numbers of identified proteins across the three superkingdoms.

Across the 100 organisms, we identified 349,164 proteins that were distinguishable by their identified peptides (Supplementary Table 2). These protein groups covered 1,136,558 entries, 93% of which were from TrEMBL—the section of the UniProt database (https://www.uniprot.org) that contains protein sequences predicted from genomes[17] (Fig. 1 and Extended Data Fig. 5). Because we have statistically significant evidence for the existence and correctness of our MS-derived peptide sequences, our data greatly increase the number of experimentally verified proteins, especially in bacteria and archaea. Contrary to our expectations, even well-studied model organisms still contributed many previously unknown proteins. The current Swiss-Prot database (version 2019_03, reviewed section of UniProt; see Methods) encompasses 559,634 experimentally verified proteins from all species. After taking into account proteins that have been described previously in the PRIDE/ProteomeXchange repository (https://www.ebi.ac.uk/pride/archive/), our additional 803,686 proteins more than double the number of proteins with experimental evidence.

To check the depth of proteome coverage, we inspected identifications for model organisms. With more than 5,000 identified protein groups in the yeast *Saccharomyces cerevisiae*, 9,000 in the zebrafish *Danio rerio* and 11,000 in the cotton plant *Gossypium hirsutum*, we obtained an even higher depth in comparison to previous large-scale efforts that focused on individual organisms. In prokaryotes we identified about half of all predicted genes at the protein level, representing a large fraction of the total proteome expressed in a single condition. However, this is less than the coverage obtained in several dedicated studies that used fractionation in these organisms and investigated different conditions. Eukaryotes generally have larger genomes and we identified correspondingly higher numbers of proteins (Fig. 1a). For instance, in a single human cell line, we identified 9,500 protein groups in our standardized workflow—a large proportion of the expressed proteome[6]—whereas 14 cell lines yielded 12,005 protein groups (Supplementary Table 4). Several species had very low proteome coverages. As the MS data were of similar quality in most of these cases (Supplementary Table 5), but the identification rates were low, we attribute

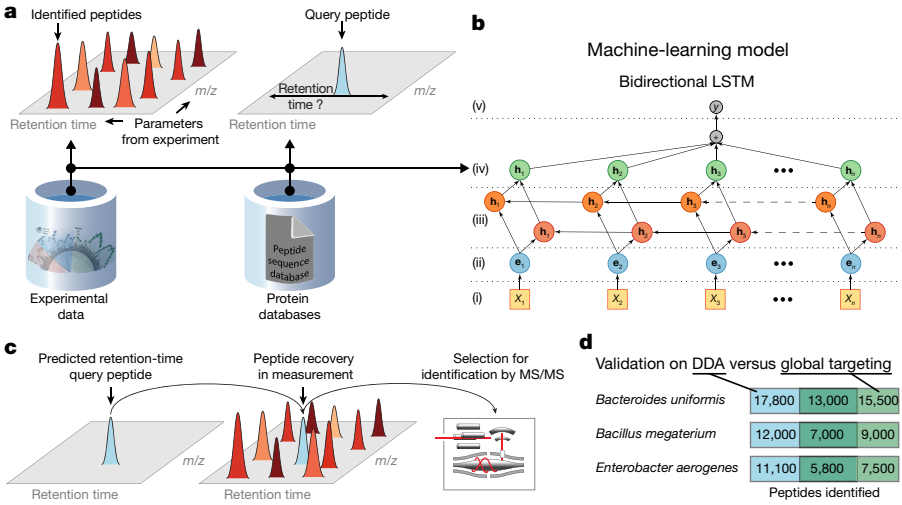


Fig. 2 | Application of a deep learning model to predict peptide retention times for liquid chromatography with tandem mass spectrometry (LC-MS/MS) measurements. a, The data used as inputs for retention time predictions are: left, our experimental data (from Fig. 1a), yielding retention time information on 2 million sequence-unique peptides from 100 organisms; and right, a list of query peptides with unknown retention times derived from a protein database. **b**, Bidirectional LSTM model with attention layer: (i), amino-acid sequence input (x_n); (ii), vectorization of amino-acid information for processing (yielding $\mathbf{e}_n$); (iii), generation of bidirectional LSTM layers ($\mathbf{h}_n$); (iv), attention-based reduction to fixed-length peptide-feature vector ($\mathbf{h}_n$); (v), prediction of retention time (y). **c**, Principle of the global targeting approach displayed for a single peptide: the instrument is set to select the peptide m/z peak for MS/MS identification if it is observed in a narrow retention time window predicted by deep learning. **d**, Application of the 'blind global targeting procedure' to all peptides of three previously unanalysed organisms resulted in the successful detection of predicted peptides in the organism samples. DDA, data-dependent acquisition.

the low proteome coverage to poor genome annotation or proteome prediction, which our data could help to improve through proteogenomics approaches.

In contrast to genomics and transcriptomics, proteomics data allow the direct estimation of the end product of gene expression[18]. We used label-free quantification in MaxQuant to estimate fractional protein intensities across multiple species[19]. Next, we asked how the proteins are distributed across the abundance range of the different organisms, and calculated the number of proteins that contribute to 90% of the total protein amount. The average was 1,546 proteins in eukaryotes, 306 in bacteria and 262 in archaea (Fig. 3a and Extended Data Figs. 6, 7). We used protein homology to enable the quantitative comparison of protein levels between the different organisms. Homology inference is a challenging bioinformatics problem, especially in poorly annotated organisms[20]. To perform the comparison across the studied species, we used high-quality homology prediction from Evolutionary Genealogy of Genes: Non-Supervised Orthologous Groups (EggNOG 5.0)[21]—a database of orthologous groups and functional annotations. We connected our quantitatively determined proteins and corresponding peptides with annotation and structural information data from various sources[17,22–24] in a graph database[25] yielding an explorable network structure with more than 8 million nodes (from proteins, peptides, gene ontology terms, and so on) and more than 53.8 million relationships between them (from homologies, associations, and so on) (Fig. 3b). The graph can be easily queried for any relationship between all of these nodes, as visualized for MS-identified homologues of two species (Fig. 3b). Here an abundant but uncharacterized protein from soybean (*Glycine max*) is linked to its counterpart in wine (*Vitis vinifera*), allowing direct comparison of MS identification, quantification and functional annotations. Similar queries can be performed for entire MS-characterized pathways, organelles or cell compartments. Co-varying pathways or gene ontology terms can also be explored, as well as their relationships to uncharacterized proteins (see www.proteomesoflife.org).

For instance, in soybean, the 11,208 quantified proteins covered more than five orders of magnitude (Fig. 3c) and had 1,763 annotated gene ontology terms. Applying a one-dimensional enrichment analysis to the annotated proteins[26] resulted in 734 statistically significantly enriched terms ($P < 0.05$) (Fig. 3d). Proteins linked to oxidation and reduction processes were the most abundant, reflecting the dominant roles of redox chemistry as a foundation for biochemical reactions such as glycolytic and carbohydrate metabolic processes (among the next most abundant categories). Apart from 'translation process', the most abundant gene ontology term of a biological process was 'protein folding', with an entire 3% of the protein mass. Altogether, functions dedicated to the life cycle of the proteome (translation, elongation, folding and proteolysis) made up a remarkable 10% of proteome mass in living organisms.

Conversely, certain classes of proteins were predominant only in specific branches of life (Extended Data Fig. 8). As expected, photosynthesis-related proteins were present only in photoautotrophic organisms such as plants, algae, protozoa or cyanobacteria (13 out of the 100 organisms) (Fig. 4 and Extended Data Fig. 9). Likewise, numerous functional associations can only be found within Bilateria or even Amniota. These mainly concern proteins associated with differentiation and tissue formation, higher intracellular spatial organization and well-described but subtaxonomy-specific signalling cascades. As expected, protein phosphorylation is predominantly but not exclusively present in eukaryotes. The bacteria and archaea both encompass organisms using this process (for instance in phosphorelay signalling), yet the proportion of the proteome mass involved in it is an order of magnitude lower in these organisms than in eukaryotes.

Much of proteome regulation is accomplished by post-translational modifications, which are typically investigated using specific enrichment protocols followed by MS analysis. However, even our nonenriched workflow in combination with the pFind tool[27] yielded a very large number of peptides with post-translational modifications for which the numbers of modified peptides were proportional to the size of the identified proteome (Extended Data Fig. 10). For instance, we found 29,426 serine phosphorylation sites, almost exclusively in eukaryotes, and 2,862 phosphotyrosine sites were largely restricted to ophistokonts (Supplementary Table 3).

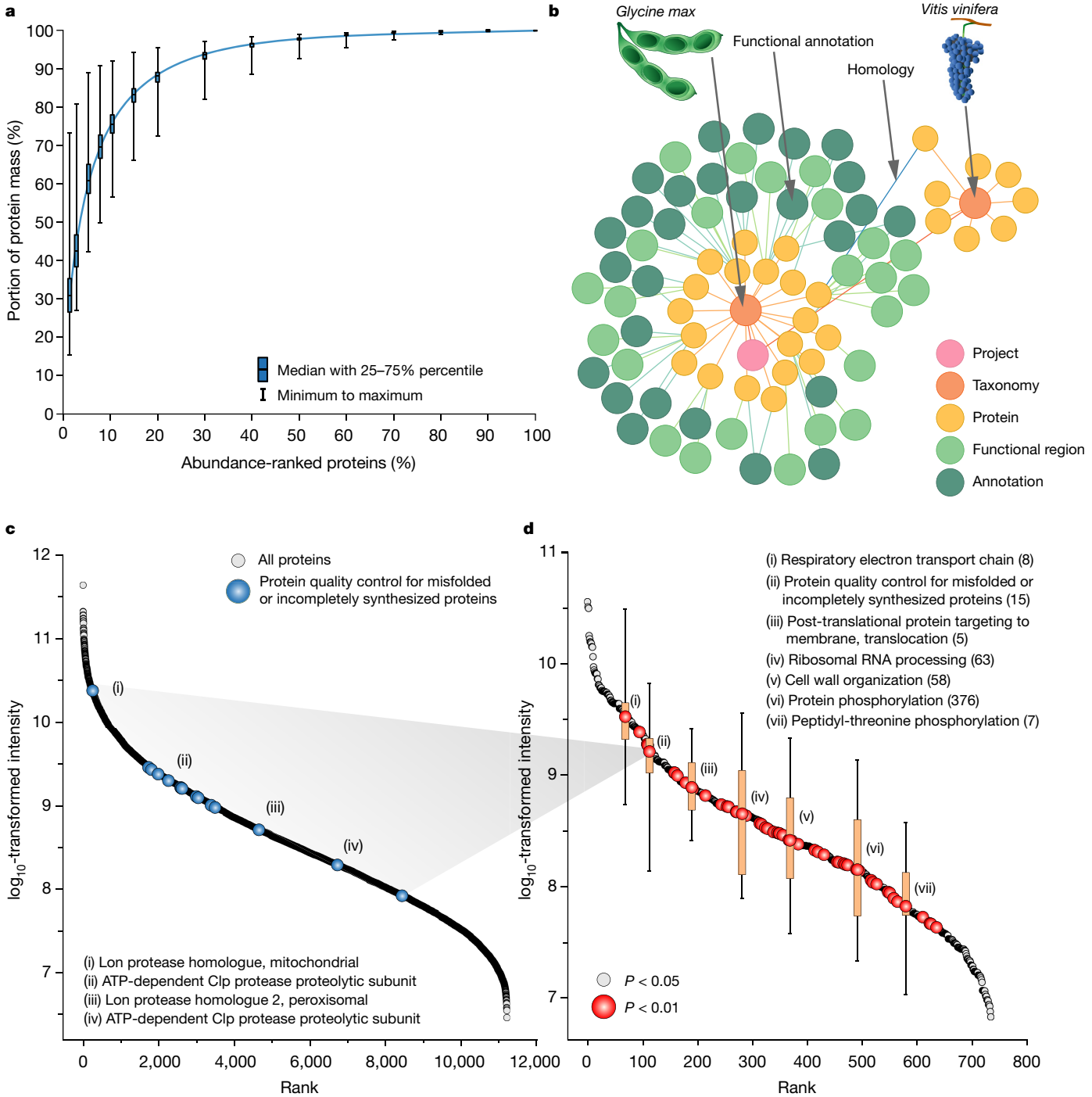


Fig. 3 | Organism-resolved integration of proteome data into a global analysis. a, Cumulative protein intensities (ranked by abundance; *x* axis) and their contribution to total protein mass (*y* axis) across all organisms ($n = 100$ organisms). **b**, Exemplified structure from the data model of the graph database, illustrating the connection between two homologous proteins of *G. max* and *V. vinifera*, and related annotations. **c**, All quantified proteins from *G. max* are displayed, plotting their intensities against their rank in the dynamic range. All proteins for which the functions are associated with 'protein quality control for misfolded or incomplete synthesized proteins' are highlighted. **d**, Significantly enriched functions (grey circles, $P < 0.05$; red circles, $P < 0.01$) within the proteome of *G. max* (with seven specific examples) and their distribution across the dynamic range (sample sizes in parentheses; one-sided Mann–Whitney *U*-test to the mean functional expression level). Error bars represent minimum to maximum values, and boxes show 10–90% percentiles.

Overall, 38.4% of the identified proteins did not have any functional annotation for the biological processes, and interestingly this was true even for 22.9% of the 100 most highly abundant proteins of each species at the biological-process level, and for 10% when considering protein functional domains (Extended Data Fig. 7 and Supplementary Table 6). Thus, our data point to a very large number of highly expressed proteins without any functional annotation or sequence homology to proteins with known gene ontology terms. Exploration of this part of the 'dark proteome' would be attractive: these proteins may indicate essential but unique features in the evolutionary development of these organisms that may be of biological or biotechnological interest.

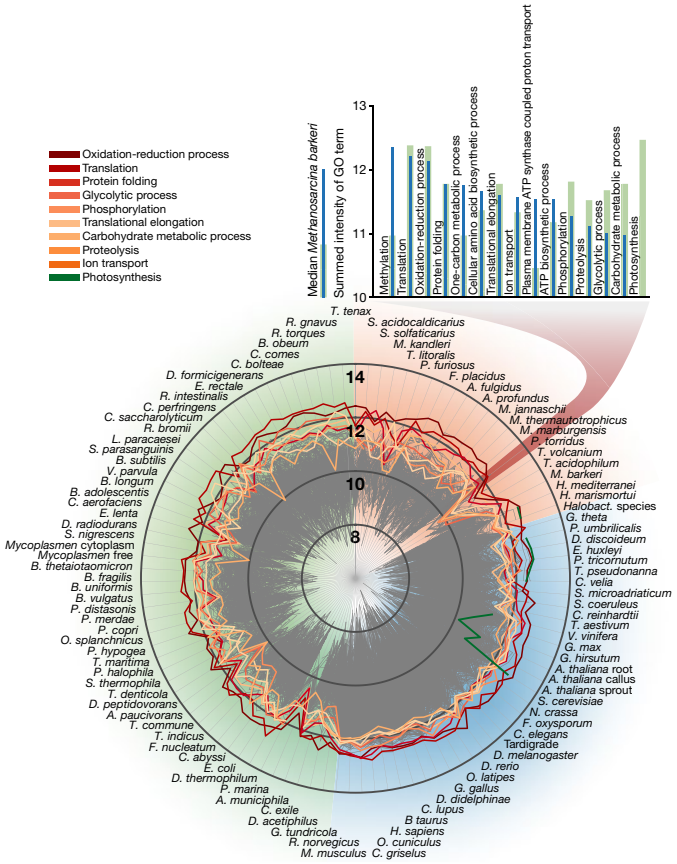


Fig. 4 | Global view of the expression levels of functional groups across the 100 organisms. The main diagram shows summed intensities for functional terms (grey lines), with the ten most abundant terms in all organisms colour-coded according to the key in the top left. The inset in the top right shows the most abundant gene ontology (GO) terms for the archaea *Methanosarcina barkeri* (blue lines), together with the median abundance of all 100 organisms for the displayed terms (green lines).

Advances in sequencing technology are now delivering the genome sequences of an exponentially increasing number of organisms, and we here made a first step towards a parallel scale-up of the characterization of proteomes. Sampling across the taxonomy of life, we created a large set of proteomes with high coverage of their expressed proteins. Label-free quantification values allow us to infer common and specialized biological functions and to compare them to close and distant relatives from all taxonomic levels. The data can be interactively explored at www.proteomesoflife.org.

Limitations of this study include the fact that we measured only selected cell types, tissues and biological states, and that the depth of proteome coverage is not yet comprehensive. Likewise, we have hardly touched upon the post-translational modification of proteins and their evolutionary diversity[28]. Ongoing improvements in MS-based proteomics—including more-refined abundance estimates[29], as well as entire streamlined workflows as described here—will substantially increase throughput in the future[2]. Given the cost effectiveness of proteomic measurements (marginal costs of less than $1,000 per species if its genome is available) and considering the wealth of novel data generated, we propose a community effort to explore many more organisms in different functional states. Integration with genomic, metabolomic and other data, together with incorporation of machine learning methods for species-specific libraries, would expand the systems-biological perspective beyond model organisms to the entire tree of life.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2402-x.

1. de Godoy, L. M. F. et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
2. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
3. Nagaraj, N. et al. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11**, M111.013722 (2012).
4. Kim, M.-S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
5. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
6. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
7. Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J. & von Mering, C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **10**, 1297–1306 (2010).
8. Marx, H. et al. A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat. Biotechnol.* **34**, 1198–1205 (2016).
9. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017); correction *Nature* **568**, E11 (2019).
10. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
11. Geyer, P. E. et al. Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2**, 185–195 (2016).
12. De Beeck, J. O. et al. Digging deeper into the human proteome: a novel nanoflow LCMS setup using micro pillar array columns (μPAC™). Preprint at *bioRxiv* https://doi.org/10.1101/472134 (2018).
13. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
14. Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
15. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
16. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
17. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47** (D1), D506–D515 (2019).
18. Muñoz, J. & Heck, A. J. R. From the human genome to the human proteome. *Angew. Chem. Int. Edn* **53**, 10864–10866 (2014).
19. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
20. Altenhoff, A. M. et al. Standardized benchmarking in the quest for orthologs. *Nat. Methods* **13**, 425–430 (2016).
21. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47** (D1), D309–D314 (2019).
22. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47** (D1), D330–D338 (2019).
23. Geer, L. Y. et al. The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492–D496 (2010).
24. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47** (D1), D427–D432 (2019).
25. Santos, A. et al. Clinical knowledge graph integrates proteomics data into clinical decision-making. Preprint at bioRxiv https://doi.org/10.1101/2020.05.09.084897 (2020).
26. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13** (Suppl 16), S12 (2012).
27. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
28. Zielinska, D. F., Gnad, F., Schropp, K., Wiśniewski, J. R. & Mann, M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol. Cell* **46**, 542–548 (2012).
29. Wiśniewski, J. R., Wegler, C. & Artursson, P. Multiple-enzyme-digestion strategy improves accuracy and sensitivity of label- and standard-free absolute quantification to a level that is achievable by analysis with stable isotope-labeled standard spiking. *J. Proteome Res.* **18**, 217–224 (2019).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample preparation

Organisms were obtained as stated in Supplementary Table 1. Cell lines were implicitly authenticated by MS and tested for mycoplasma contamination. The LLC-PK1 cell line was contaminated and mycoplasma contamination was harvested for analysis.

We carried out sample preparation according to the in-StageTip protocol[10] with an automated set-up on an Agilent Bravo liquid-handling platform as described[11]. In brief, samples were incubated in PreOmics lysis buffer (catalogue number P.O. 00001, PreOmics) for reduction of disulfide bridges, cysteine alkylation and protein denaturation at 95 °C for 10 min. Root and sprout parts of *Arabidopsis thaliana*, whole *Drosophila melanogaster* and leaves of *Porphyra umbilicalis* were ground in liquid nitrogen with a mortar and pestle beforehand. Samples were sonicated using a Bioruptor Plus from Diagenode (15 cycles, each of 30 s), and the protein concentration was measured using a tryptophan assay. In total, 200 µg of protein from each organism were further processed on the Agilent Bravo liquid-handling system by adding trypsin and LysC (at a 1:100 ratio of enzyme to sample protein, both in micrograms), mixing and incubating at 37 °C for 4 h.

We purified the peptides in consecutive steps according to the PreOmics iST protocol (www.preomics.com). After elution from the solid-phase extraction material, the peptides were completely dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator Plus). Peptides were suspended in buffer A* (2% acetonitrile (v/v), 0.1% trifluoroacetic acid (v/v)) and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Eukaryotes generally have larger numbers of genes than bacteria and archaea, resulting in a larger number of proteins and consequently of peptides. To reduce the complexity in the MS measurements, we separated eukaryotic peptide mixtures into eight fractions using the high-pH reversed-phase 'spider fractionator' as described[13].

UHPLC and mass spectrometry

We analysed the samples by applying LC-MS instrumentation, comprising an EASY-nLC 1200 ultrahigh-pressure system (Thermo Fisher Scientific) coupled to a Q Exactive HFX Orbitrap instrument[30] (Thermo Fisher Scientific) with a nano-electrospray ion source (Thermo Fisher Scientific).

For each analysis, 500 ng of purified peptides were separated on a 200 cm µPAC C_{18} microchip nano-LC column (PharmaFluidics). Peptides were loaded in buffer A*. To overcome the void volume of 10 µl, we applied a concentration gradient from 5% buffer B (0.1% formic acid (v/v), 80% acetonitrile (v/v)) to 10% buffer B coupled with a flow gradient from 750 nl min^{-1} to 300 nl min^{-1} for the first 15 min. Subsequently peptides were eluted with a linear gradient from 10% to 30% buffer B in 125 min at a constant flow rate of 300 nl min^{-1}. This was followed by a stepwise increase of buffer B to 60% in 5 min and to 95% buffer B in 5 min. Afterwards we applied a 5 min wash with 95% buffer B, followed by a 5 min decrease to 1% buffer B and a 20 min wash. We kept the column temperature constant at 50 °C by using an oven from Phoenix S&T (catalogue number PST-BPH-15). To avoid interference between the electrospray voltage and the µPAC chip column, we grounded the post-column connection, which was connected by a 20 cm long, 20 µm inner diameter fused silica post-column line to a New Objective Pico-Tip Emitter. This setup is further detailed in Extended Data Fig. 1b. The electrospray voltage was applied by connecting the mass spectrometer source output to the metal connection between the post-column sample line with an in-house-made clamp connection.

HPLC parameters were monitored in real time using SprayQC software[31]. MS data were acquired with a Top15 data-dependent MS/MS method. Target values for the full-scan MS spectra were 3×10^6 charges in the m/z range 300–1,650, with a maximum injection time of 20 ms and a resolution of 60,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27 eV. MS/MS scans were performed at a resolution of 15,000 at m/z 200 with a target value of 1×10^5 and a maximum injection time of 28 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data analysis

MS raw files were analysed using MaxQuant software, version 1.6.1.13 (ref. [32]), and peptide lists were searched against their species-level UniProt FASTA databases. A contaminant database generated by the Andromeda search engine[33] was configured with cysteine carbamidomethylation as a fixed modification and amino-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels, with a minimum length of seven amino acids for peptides. The FDR was determined by searching a reverse database. Enzyme specificity was set as carboxy-terminal to arginine and lysine as expected, using trypsin and LysC as proteases. A maximum of two missed cleavages was allowed. Peptide identification was performed in Andromeda with an initial precursor mass deviation of up to 7 ppm and a fragment mass deviation of 20 ppm. All proteins and peptides matching the reversed database were filtered out. All bioinformatics analyses were performed using Perseus[34] as well as standard analysis in Python version 3.6.4.

Machine learning model to predict retention times

To predict the retention times of peptides by machine learning, we isolated all detected peptide sequences, including modified peptides. For solvent-induced microshifts between runs, we corrected the detected retention times per peptide by the median shift of all peptides from one run to the median peptide retention time. This resulted in a total of 5,168,800 peptide sequences corresponding to 2,196,869 unique peptide sequences with a median retention time value for retention time prediction.

Our neural network architecture model takes a raw peptide sequence as input. Each amino acid was encoded into a 26-dimensional vector representation for processing using a one-hot encoding scheme, resulting in an $Lx26$ feature vector for a peptide with length L. This vector was connected to a two-layer bidirectional recurrent network with LSTM units with 500 hidden nodes each, which extract context-based features for each individual amino acid. This amino-acid-based feature embedding was reduced to a global 128-dimensional peptide-feature vector by an attention layer, which predicts the contribution of each individual amino-acid feature vector to the regression task. This peptide-feature vector was the input to a logistic regression layer, which regresses the expected retention time for the peptide sequence. The combination of recurrent layers with the attention layer allowed the model architecture to process peptide sequences with arbitrary lengths, but at the same time allow interpretability. The model was end-to-end trained on 2,125,113 peptides and validated on 54,490 holdout peptides. To validate the retention time prediction in vitro, we used the trained model to predict the peptide retention times of all tryptic peptides from *B. uniformis*, which were not included in the training set. We set the mass spectrometer to sequence only if the peptide eluted in a window of 1.4 s around the predicted retention time. This 'global targeting' was done using MaxQuant.life software (version 0.15)[35].

Graph database and cloud data-analysis notebook

To allow exploration of the MS experimental results, we developed a graph database (Neo4j: http://neo4j.com/, version 3.5.8, community edition) that collects all of the experimental data as well as homology and

functional annotations from different publicly available resources[17,21–24,36]. The implemented data model contains 11 different types of node and 14 types of link among the nodes; the data amount to 7,410,594 nodes and 35,517,979 relationships (5.02 GB). To populate the graph, flat files from source databases were downloaded and parsed to generate tab-delimited files comprising nodes and relationships, and standardized using selected terminologies and ontologies. The relationships collected in the database describe ontology structures (Directed Acyclic Graph relationships) and homology (orthology or paralogy) or functional associations (biological processes, functional regions, and so on). A version of the database is accessible at http://www.proteomesoflife.org.

The website gives access to interactive analyses implemented in Python (version 3.6), and uses Cypher as the query language (https://neo4j.com/developer/cypher-query-language/) (see also ref. [37]).

Data integration and comparison

We compared data in online proteomics repositories (PRIDE (https://www.ebi.ac.uk/pride/) and ProteomeXchange (http://www.proteomexchange.org)) with our data from 100 organisms, and downloaded either the provided protein tables or the raw files (Supplementary Table 6). We analysed the raw files with the same MaxQuant version and sequence files as used in our study. If identifiers other than UniProt identifiers were used, we applied the UniProt database to find the corresponding entries and to determine those proteins for which there was previous MS evidence.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The MS-based proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with identifier PXD014877 and PXD019483.

Code availability

Custom computer code is available at https://github.com/MannLabs/proteomesoflife.

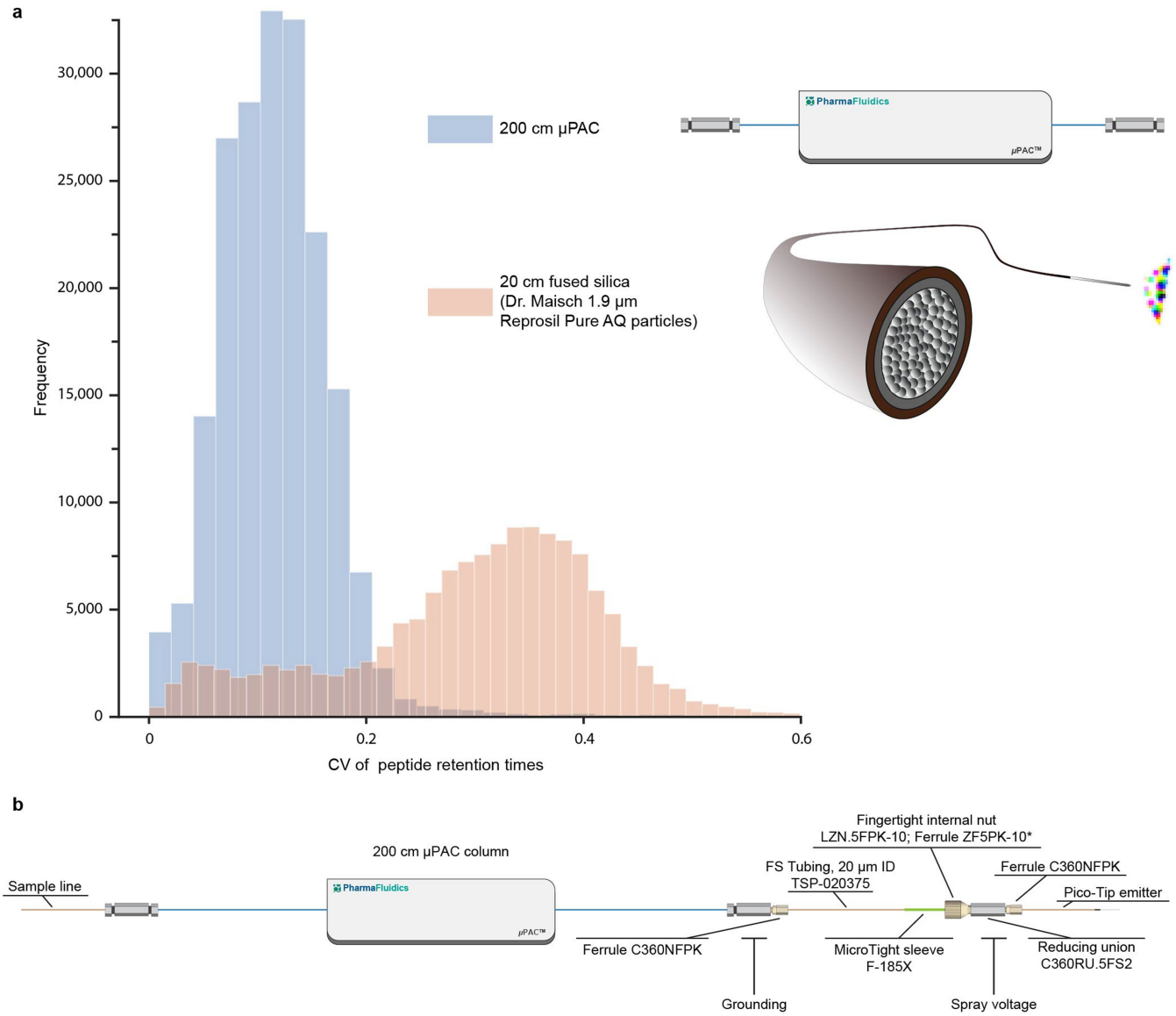
30. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
31. Scheltema, R. A. & Mann, M. SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11**, 3458–3466 (2012).
32. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
33. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
34. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
35. Wichmann, C. et al. MaxQuant.Live enables global targeting of more than 25,000 peptides. *Mol. Cell. Proteomics* **18**, 982–994 (2019).
36. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47** (D1), D442–D450 (2019).
37. Perkel, J. M. Why Jupyter is data scientists' computational notebook of choice. *Nature* **563**, 145–146 (2018).

Acknowledgements We thank all members of the Proteomics and Signal Transduction Group and the Clinical Proteomics Group at the Max Planck Institute of Biochemistry, Martinsried, for help and discussions, and in particular I. Paron, C. Deiml, A. Strasser and B. Splettstoesser for technical assistance. We further thank the P. Bork group for supplying bacteria, the A. Pichlmair group for virus samples, F. Hosp for *A. thaliana*, I. Sinning for *Neurospora crassa* and the K.-P. Janssen group for cell line samples. Our work was partially supported by the Max Planck Society for the Advancement of Science, by the European Union's Horizon 2020 research and innovation program with the Microb-Predict project (grant 825694), by grants from the Novo Nordisk Foundation (NNF15CC0001 and NNF15OC0016692), and by the Deutsche Forschungsgemeinschaft (DFG) project 'Chemical proteomics inside us' (grant 412136960).

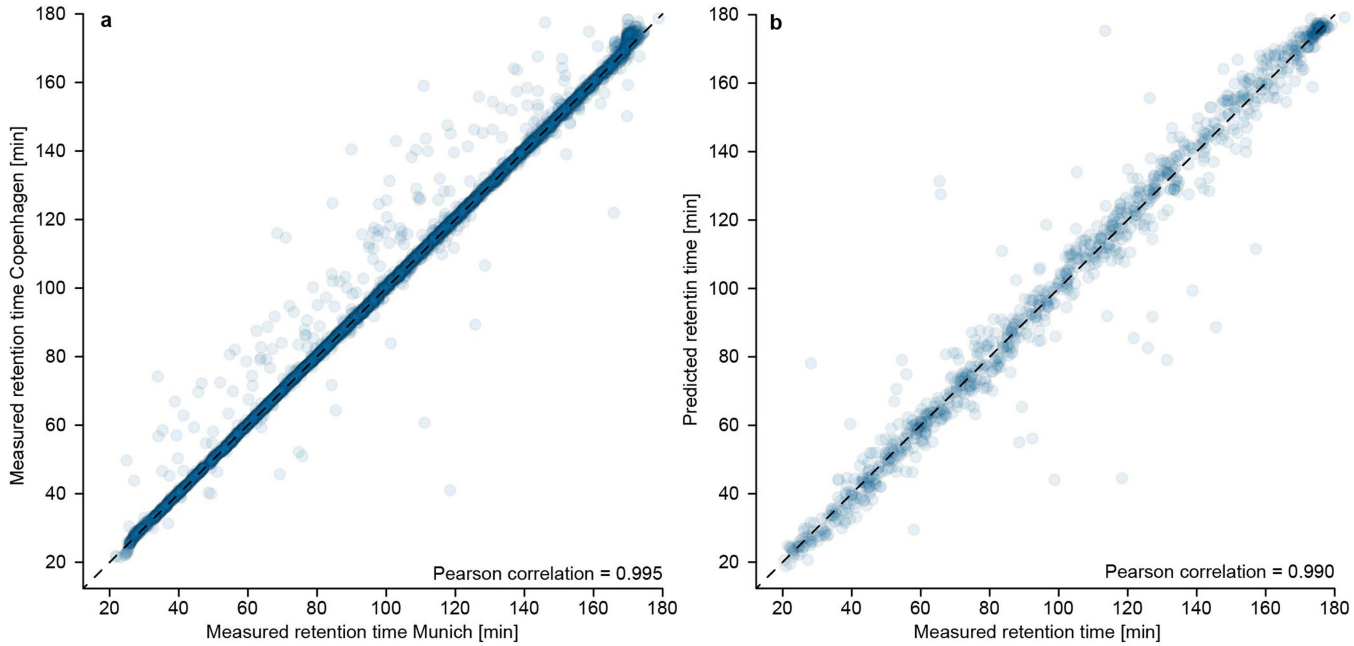
Author contributions J.B.M. and P.E.G. designed the experiments, performed and interpreted the MS-based proteomic analyses, carried out bioinformatics analyses and generated text and figures for the manuscript. P.V.T., S.D., S.V.W. and J.M.B. designed experiments and performed MS-based proteomics analyses. A.R.C. and A.S. integrated annotation data with proteomics data and implemented the Python code as well as graph-based structures. A.S. and M.O. implemented the web-accessible analyses. N.K., F.T. and M.T.S. carried out the machine learning analysis. M.M. supervised and guided the project, designed the experiments, interpreted MS-based proteomics data and wrote the manuscript.

Competing interests The authors declare no competing interests.

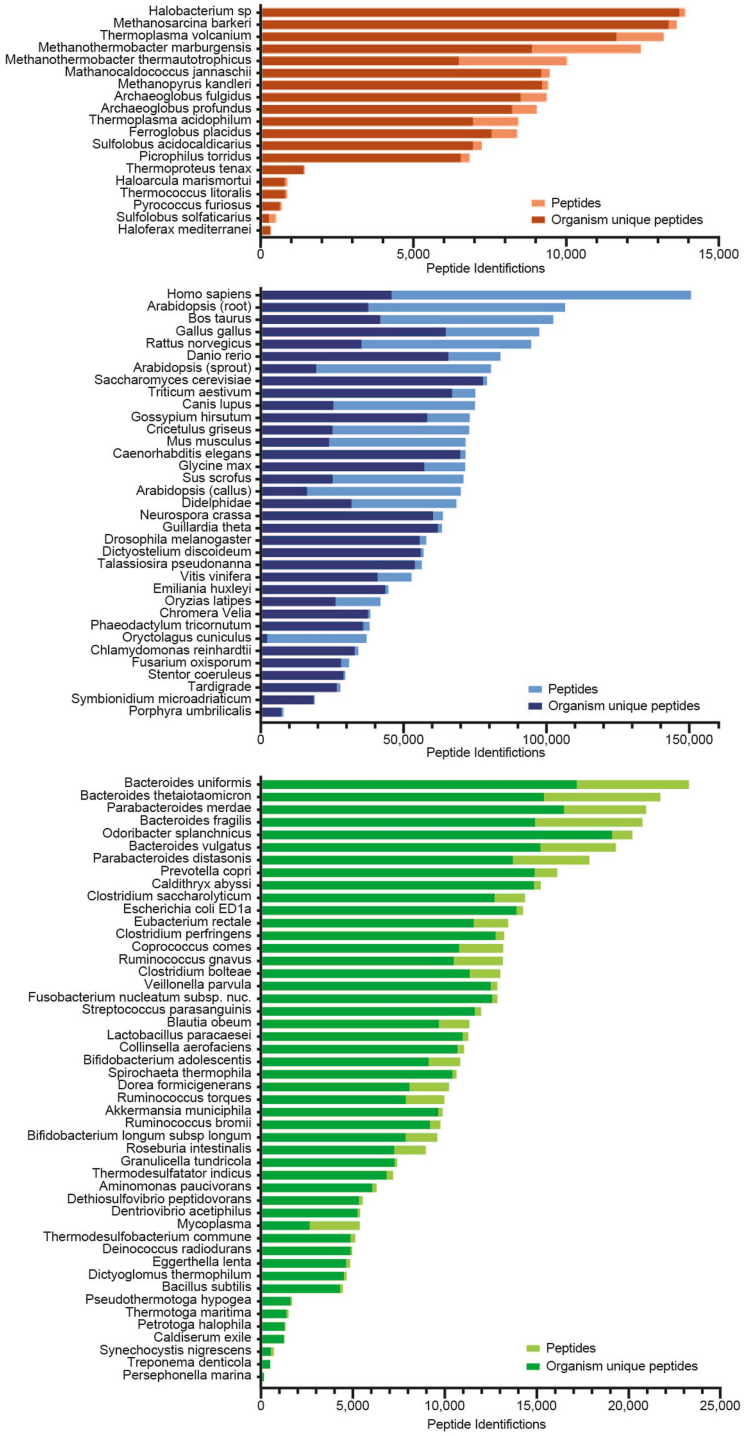
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2402-x.
Correspondence and requests for materials should be addressed to M.M.
Peer review information *Nature* thanks Joshua Coon, Vera van Noort and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Comparison of the peptide retention times obtained by a µPAC and a fused silica capillary column. a, The histograms illustrate the distribution of coefficients of variation (CVs) calculated from peptide retention times obtained by a µPAC and a fused silica capillary column. The CVs were calculated for peptides from 12 measurements of a HeLa cell digest on each column. **b**, All components, including lines, connectors, the column and the emitter, are displayed together with grounding and spray voltage connections. The pico tip emitter is from New Objective (catalogue number FS360-20-10-N-5-105CT).

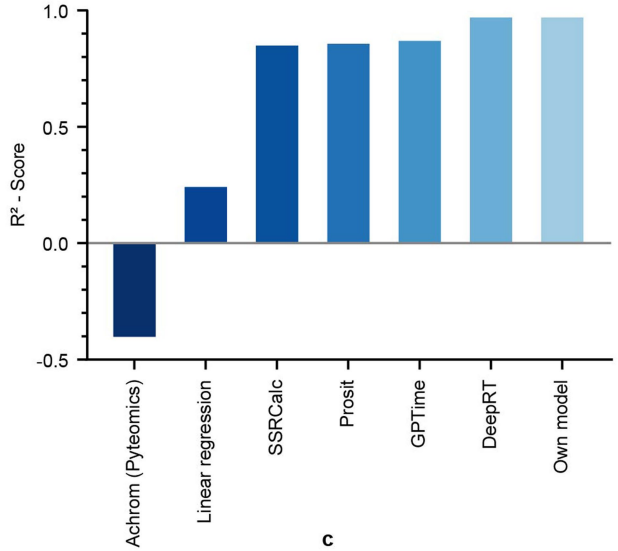


Extended Data Fig. 2 | Interlaboratory reproducibility and prediction of peptide retention time on the µPAC column. a, The ability to produce chip-based columns in a reproducible manner, coupled with the statically fixed micrometre-sized pillars, results in highly reproducible performance and interlaboratory transferability of the µPAC-based approach. Shown are the corrected retention times of an excerpt of 5,000 peptides from the 43,000 overlapping peptides measured in two different HeLa cell digests by our Munich and Copenhagen laboratories, resulting in a Pearson correlation coefficient of peptide retention times of 0.995. **b**, To validate our model for predicting peptide retention times, we plot an excerpt of 1,000 peptides from the complete test-set of 54,490 peptides, with experimentally determined values on the *x* axis and predicted values on the *y* axis. The Pearson's R^2 correlation value for the complete predicted peptide set is 0.99.



Extended Data Fig. 3 | Total numbers of identified peptides from 100 organisms across the tree of life. The peptides uniquely identified for a certain organism are colour-coded from peptides identified in multiple species. Orange, archaea; blue, eukaryotes; green, bacteria.

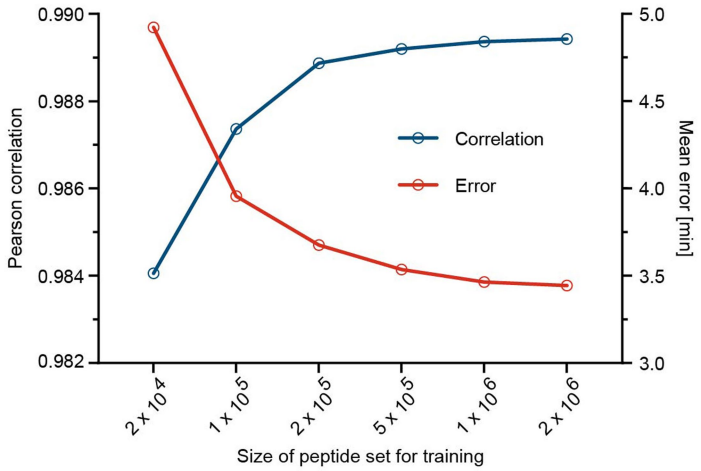
a



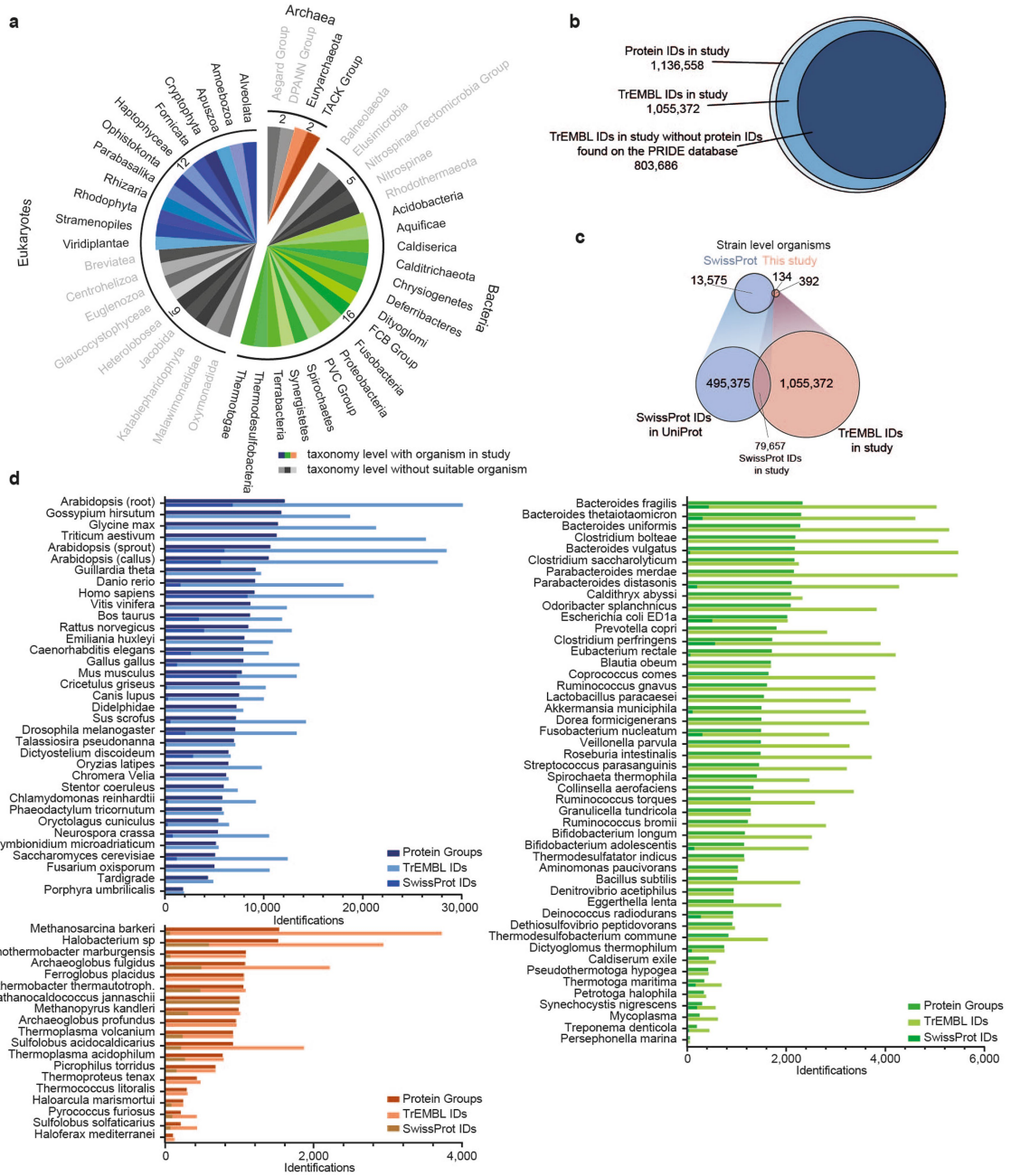
b

Model	R2-Score	Pearson	Training set	Test set
Achrom (Pyteomics)	-0.413	0.232	10000	10000
Linear Regression	0.253	0.503	1537808	659060
SSRCalc	0.860	0.928	1000	1000
Prosit	0.868	0.969	-	1000
GPTime	0.881	0.939	10000	10000
DeepRT	0.980	0.990	1537808	659060
Own model	0.980	0.990	2125113	50000

c

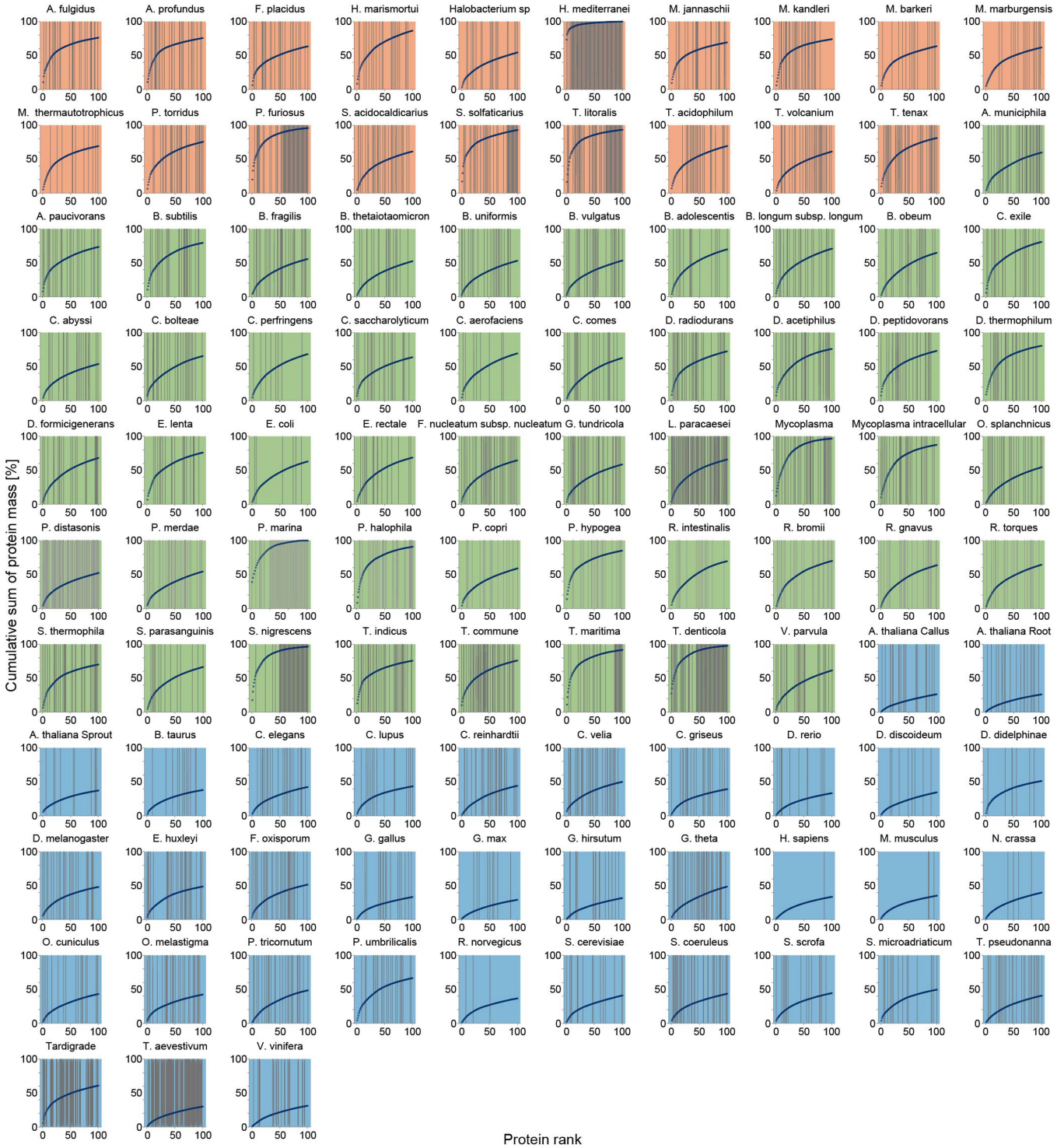


Extended Data Fig. 4 | Comparison and characterization of the LSTM model for predicting peptide retention times. a, Box plots comparing R^2 scores obtained from different models of peptide retention time, calculated from the linear regressions of correlations between the predicted test set to the measured peptide retention times. Sample sizes are shown in **b**. **b**, Table comparing the different models of peptide retention time. The training set was reduced in size (number of peptides included) in order to account for the exponentially growing calculation time of certain models. Statistics represent the linear regression of correlation from the predicted test set retention times to the measured retention times. **c**, Characterization of the LSTM model applied here for different sizes of training peptide set.

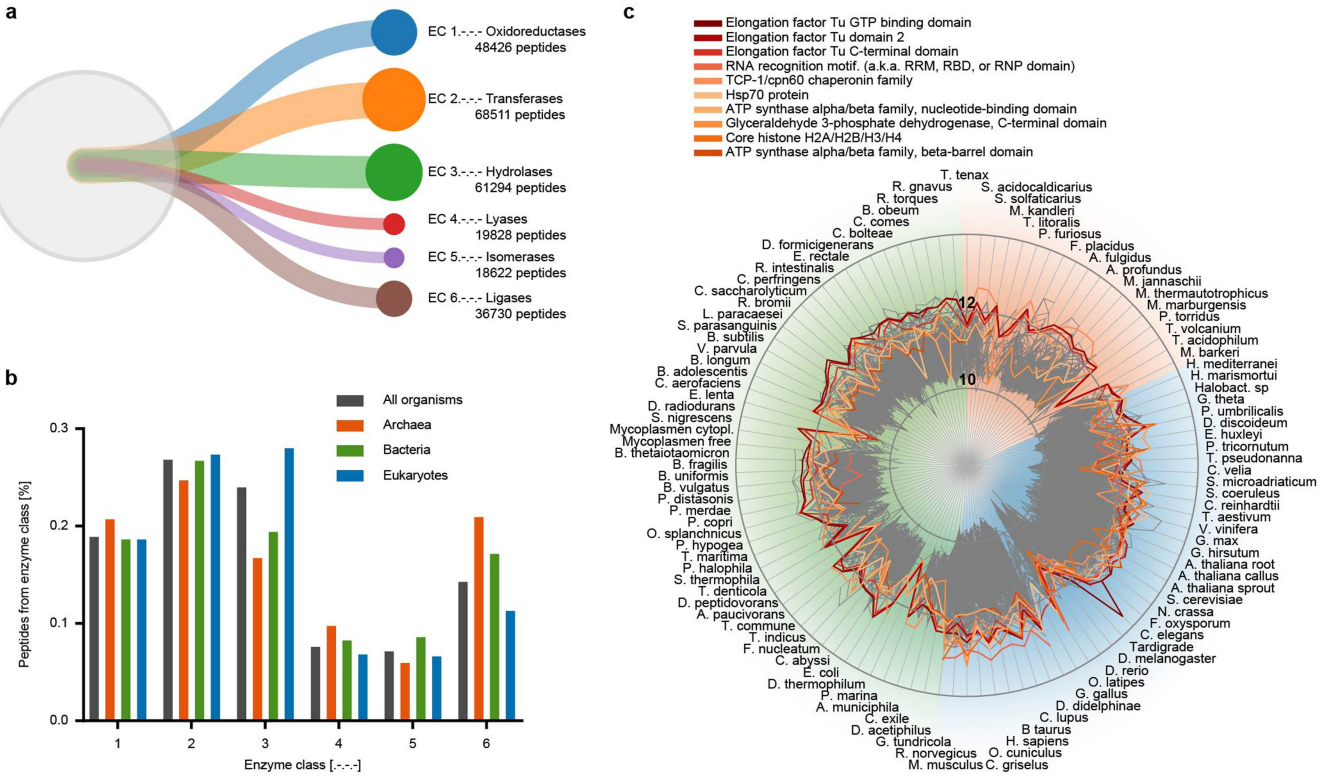


Extended Data Fig. 5 | Overview of our data set of 100 organisms across the tree of life. a, Illustration of all direct taxonomic levels below the superkingdom level that are covered by our data set. DPANN, Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota; FCB, Fibrobacteres, Chlorobi and Bacteroidetes; PCV, Planctomycetes, Chlamydiae and Verrucomicrobia; TACK, Thaumarchaeota, Crenarchaeota and Korarchaeota. **b**, Number of protein identification codes (IDs) in this study and their relation to TrEMBL IDs found in the PRIDE archive. **c**, Comparison of the Swiss-Prot database to the data set in this study with regards to organism and protein numbers. **d**, Numbers of identified protein groups and UniProt protein entries for all 100 organisms in our data set. The UniProt protein-entry identifications are colour-coded into Swiss-Prot (reviewed) and TrEMBL (predicted) entries.

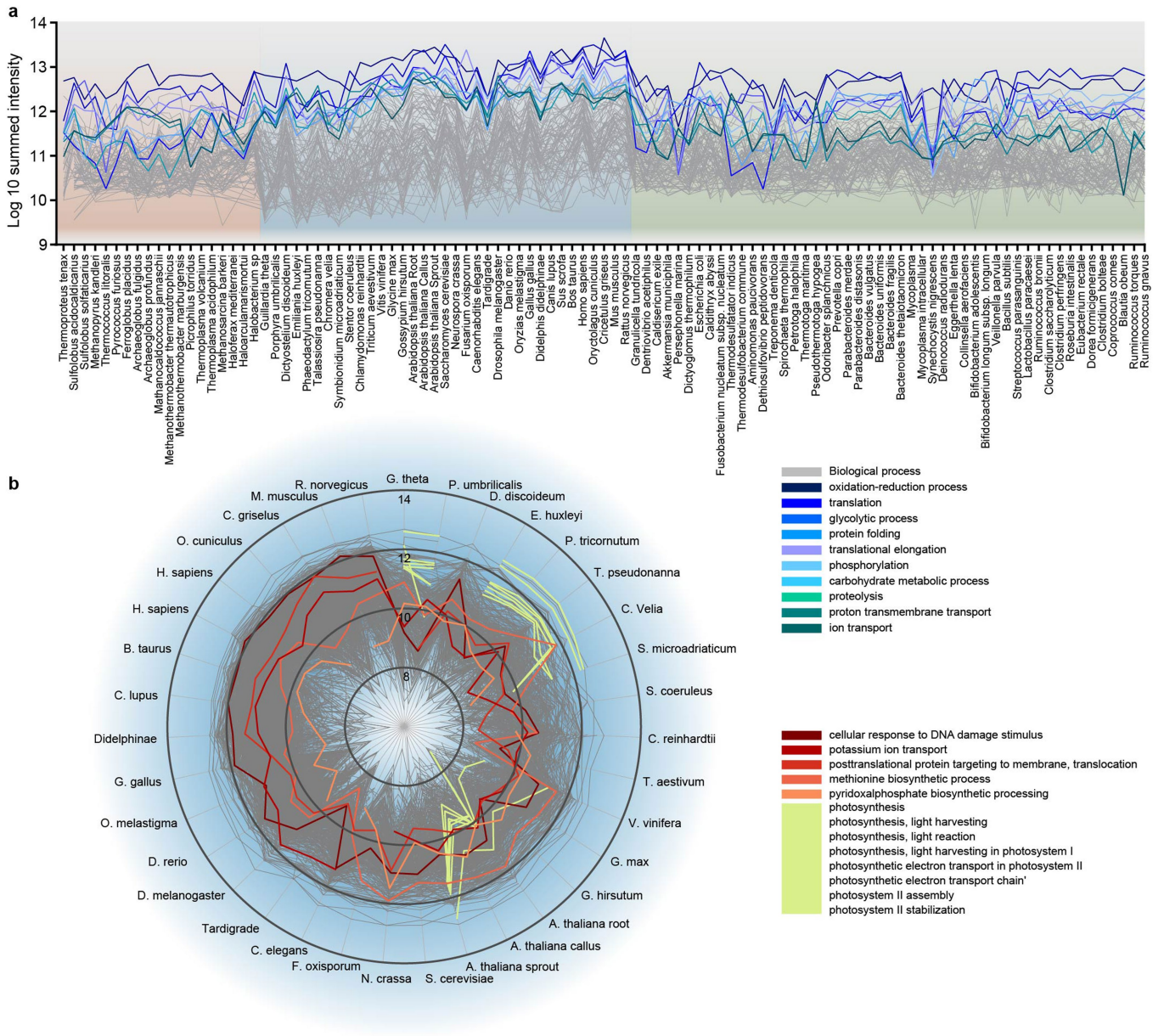
Extended Data Fig. 6 | Dynamic range curves for all organisms analysed here. Protein intensities are $\log_{10}$-scaled and plotted against the abundance rank of each protein.



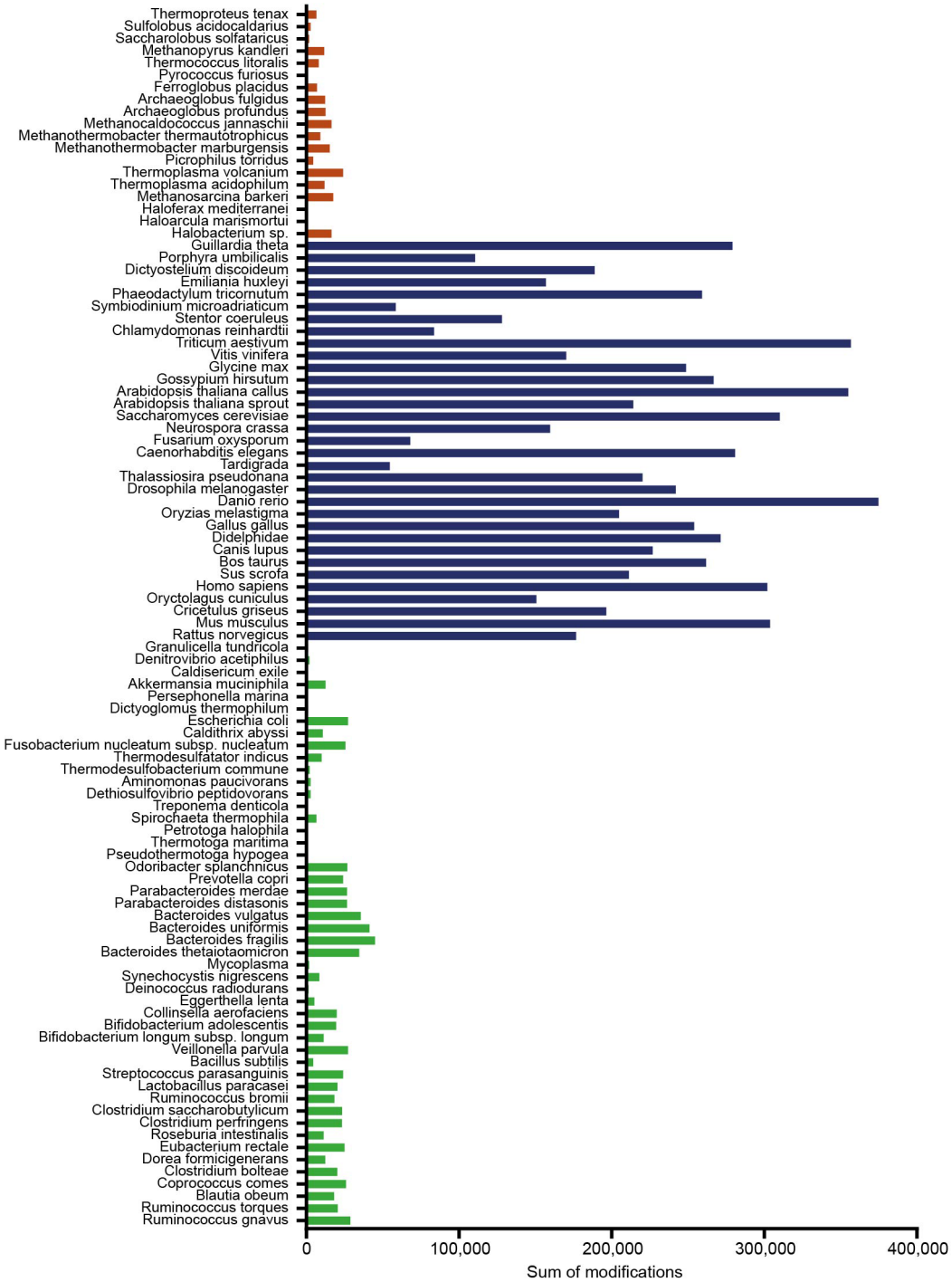
Extended Data Fig. 7 | Cumulative protein intensities for all organisms analysed here. On the *x* axis, proteins are ranked according to their abundance; the *y* axis shows the cumulative protein intensity. Proteins missing biological-process annotation are highlighted by grey lines in the background.



Extended Data Fig. 8 | Quantitative analysis of different enzyme classes and functional protein domains across the tree of life. a, We classified the contribution of peptides to the top 90% of protein mass within all 100 organisms according to the enzyme commission (EC) number, using the Unipept web-tool (https://unipept.ugent.be/). The alluvial plot illustrates the proportions of each enzyme class across all organisms in our study. **b**, Comparison of the three domains of life with respect to their normalized contribution of peptides to each enzyme class. **c**, Proteins that contribute to the top 90% of the protein mass within all 100 organisms studied herein were annotated according to their known functional protein domains, and the intensities for different functional domains of an organism were summed to display the most abundant functional protein domains across the tree of life. The intensity is displayed on a $\log_{10}$ scale.



Extended Data Fig. 9 | Quantitative analysis of specific biological processes across the tree of life. a, Linear display showing a global view of the expression levels of functional groups across the 100 organisms from Fig. 4. Summed intensities for functional terms are shown as grey lines, with the top ten most abundant terms in all organisms colour-coded according to the top key. **b**, Quantitative analysis of specific biological processes from the superkingdom of eukaryotes. Proteins were annotated with biological processes, and the intensities for each annotation term within an organisms were summed. Those biological processes that display differential expression across the superkingdom as well as photosynthetic processes are highlighted according to the bottom key.



Extended Data Fig. 10 | Modified peptides. Sum of modified peptides per organism, identified with pFind (http://pfind.ict.ac.cn/software/pFind3/index.html) and colour-coded for archaea (red), eukaryotes (blue) and bacteria (green).

nature research

Corresponding author(s): Prof Matthias Mann

Last updated by author(s): Feb 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The LC-MS data collection was performed with a Thermo Fisher Easy LC 1200, employing a Pharmafluidics 200 cm µPAC column using a Thermo Fisher Q Exactive HF-X mass spectrometer. Electrospray conditions were monitored by the SprayQC software.
Data analysis	LC-MS data processing was performed using MaxQuant software (version 1.6.1.13), employing the Andromeda search engine. Data analysis was done using Perseus software and custom code written in Python (version 3.6). For the interactive analyses a Neo4j (version 3.5.8 Community edition) was used to develope a graph database which is acces via Python (version3.6) and uses Cypher as query language.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

LC-MS raw files, FASTA files and MaxQuant output tables are available via the PRIDE partner repository via ProteomeXchange with identifier PXD014877.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	100 Organisms across the tree of life, thereof 19 archaea, 49 bacteria and 32 eukaryotes were analyzed in singlets by MS based proteomics.
Research sample	To collect a diverse set of representative organisms across the tree of life, we considered the availability of assembled genome sequences, the accessibility of cultured or tissue material and included common model organisms for comparison.
Sampling strategy	Cohort size of 100 organisms was chosen by a trade of between biological impact and data acquisition time. As a pioneer study in cross organism proteome comparison this represents a sufficient overview of the proteome of living organisms. Samples were prepared for bottom up proteomics with LC-MS, by denaturationg and lysing (boiling, sonication, grinding), reduction/ alkylation and tryptic digestion of proteins, followed by subsequent purification of peptides by desalting.
Data collection	Data was obtained by bottom up proteomic measurement of tryptic digested proteins with mass spectrometry. The experiments were carried out by the authors.
Timing and spatial scale	Data were collected from September 2018 to January 2019.
Data exclusions	No data were excluded.
Reproducibility	No repetitive measurements were conducted.
Randomization	Sample handling and measuring was done in batch processing. Fractions of the same samples were measured consecutively.
Blinding	Samples were not blinded due to batch processing.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC-CCL-22; ATCC-CCL-34; ATCC-CCL-61; ATCC-CRL-2050; ATCC-CRL-1840; 6C2; ATCC-CCL2; ATCC-CRL-3209; ATCC-CRL-6503; ATCC-CRL-1721; LLC-PK1
Authentication	Cell lines were not authenticated.
Mycoplasma contamination	One cell line was positive for Mycoplasma contamination which was used to generate a proteomics dataset for Mycoplasma.

Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Drosophila melanogaster, strain CantonS with a mean age of 24 h and mixed sex were used. Caenorhabditis elegans N2 wild.type strain were used.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	No ethical approvement was required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.