Functional Optimization in Distinct Tissues and Conditions Constrains the Rate of Protein Evolution

Dinara R. Usmanova [1,†] Germán Plata [1,2,†] Dennis Vitkup [1,3,*]

[1]Department of Systems Biology, Columbia University, New York, NY 10032, USA
[2]BiomEdit, Fishers, IN 46037, USA
[3]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[†]These authors contributed equally.

***Corresponding author:** E-mail: dv2121@cumc.columbia.edu.

Associate editor: John Parsch

Abstract

Understanding the main determinants of protein evolution is a fundamental challenge in biology. Despite many decades of active research, the molecular and cellular mechanisms underlying the substantial variability of evolutionary rates across cellular proteins are not currently well understood. It also remains unclear how protein molecular function is optimized in the context of multicellular species and why many proteins, such as enzymes, are only moderately efficient on average. Our analysis of genomics and functional datasets reveals in multiple organisms a strong inverse relationship between the optimality of protein molecular function and the rate of protein evolution. Furthermore, we find that highly expressed proteins tend to be substantially more functionally optimized. These results suggest that cellular expression costs lead to more pronounced functional optimization of abundant proteins and that the purifying selection to maintain high levels of functional optimality significantly slows protein evolution. We observe that in multicellular species both the rate of protein evolution and the degree of protein functional efficiency are primarily affected by expression in several distinct cell types and tissues, specifically, in developed neurons with upregulated synaptic processes in animals and in young and fast-growing tissues in plants. Overall, our analysis reveals how various constraints from the molecular, cellular, and species' levels of biological organization jointly affect the rate of protein evolution and the level of protein functional adaptation.

Key words: **protein evolution, molecular clock, protein function, functional optimization, expression cost.**

Discoveries

Introduction

Understanding protein molecular evolution and functional adaptation is a long-standing goal of biological research. The rate of protein evolution, i.e. the number of amino acid substitutions per protein site per unit time, remains approximately constant across different lineages (Zuckerkandl and Pauling 1962, 1965) but varies by orders of magnitude across cellular proteins (Dickerson 1971; Koonin and Wolf 2010). It is currently unclear what are the main biological mechanisms underlying this rate variability (Zhang and Yang 2015). Alongside elucidating the determinants of protein evolution, another key challenge in molecular and cell biology is understanding how and to what extent protein function is optimized in the context of different cell types and tissues. Previously, it has been observed that protein function, such as enzymatic efficiency, appears to be only moderately optimized in various species (Bar-Even et al. 2011). But the origins of this diversity in functional optimization between proteins are not understood. Although the questions concerning the variability of protein evolutionary rates and protein functional optimization have been rarely considered together, they are likely to be closely intertwined. Because the majority of newly arising mutations are harmful to protein function (Futuyma and Kirkpatrick 2017), the strength of purifying selection against deleterious mutations, and therefore the rate of protein evolution, may depend on the level of optimized protein efficiency. Although plausible, the empirical evidence for this effect and its magnitude is currently lacking.

We note that evolutionary rates vary both between different proteins and also between different amino acid sites within a protein. The variability of site-specific evolutionary rates and their long-term divergence limits are well explained by a combination of structure, stability, and function-related factors (Jack et al. 2016; Konate et al. 2019). The influence of these factors on site-specific evolutionary rates can be modeled by considering the purifying selection necessary to maintain protein stability and functional activity (Echave 2019; Ferreiro et al. 2024). In this paper, we focus on a different question, i.e. the variability of evolutionary rates between cellular proteins. Protein-specific evolutionary rates correlate only weakly with the average protein contact density (Zhou et al. 2008) and the

Received: April 18, 2024. **Revised:** July 29, 2024. **Accepted:** August 05, 2024

Open Access

overall protein stability (Plata and Vitkup 2018; Usmanova et al. 2021). This is likely explained by the fact that the proportion of sites with different biophysical properties does not vary dramatically across proteins, and that increasing protein stability beyond a certain limit is generally not advantageous to protein function (Goldstein 2011).

Multiple genomic, cellular, and molecular correlates of protein evolutionary rates have been previously considered (Koonin and Wolf 2006; Rocha 2006), and the best-known predictor is the level of protein expression (Pal et al. 2001, 2006; Rocha and Danchin 2004). The inverse relationship between protein expression and the rate of protein evolution, usually referred to as the Expression-evolutionary Rate (ER) correlation, shows that highly expressed proteins generally evolve slower than proteins with low expression; in other words, the sign of the ER correlation is negative. The level of gene expression explains up to a third of the protein evolutionary rate variance in various species, but the biological mechanisms underlying the ER correlation are not currently understood (Zhang and Yang 2015; Usmanova et al. 2021). The ER correlation in animals is usually stronger in neural tissues (Drummond and Wilke 2008; Tuller et al. 2008). However, the reasons for this interesting observation are not clear. At the molecular level, several models have been proposed to explain the ER correlation. One hypothesis suggested that ER is primarily mediated by increased protein stability necessary to prevent effects associated with toxic misfolding of highly expressed proteins (Drummond and Wilke 2008). However, multiple studies of various empirical datasets demonstrate only a small role played by protein stability in explaining the variability of protein evolutionary rates (Plata and Vitkup 2018; Biesiadecka et al. 2020; Usmanova et al. 2021; Wu et al. 2022).

The optimization of protein molecular function may not only slow the rate of protein evolution but may also underlie the aforementioned ER correlation (Rocha 2006; Cherry 2010; Gout et al. 2010). The total level of protein activity in the cell is usually proportional to the product of protein functional efficiency and protein expression level. Therefore, by optimizing protein efficiency organisms can express fewer copies of a protein while maintaining its total cellular activity. We refer to this mechanism as FORCE, as it is based on the idea of protein Functional Optimization to Reduce the Cost of Expression. According to FORCE, highly expressed proteins are under more stringent selection for functional optimization because improving their efficiency allows cells to save more resources required for protein production. Although computer simulations demonstrated how protein functional optimization, coupled with protein production costs, can in principle lead to the ER correlation (Cherry 2010), it is currently unclear whether this mechanism may explain a substantial fraction of the evolutionary rate variance across proteins and how the cost of expression in various tissues is related to protein functional optimization.

In this study, we address several interrelated scientific questions raised above using analyses of multiple functional and genomics datasets. First, based on data describing enzymes' catalytic efficiency, we investigate the role of functional optimization in constraining protein evolution. We then analyze comprehensive tissue- and cell-type-specific transcriptomics data to explore what biological and cellular processes are usually associated with strong ER correlations in animal and plant tissues. Next, we investigate the relationships between protein evolution, functional optimization, and expression patterns across tissues in multicellular organisms. Overall, our study reveals fundamental biological mechanisms underlying the variability of protein evolutionary rates and demonstrates how specific molecular and cellular processes affect protein functional optimization.

Results

Protein Evolution and Optimization of Protein Molecular Function

To explore how the selection for functional optimality influences protein evolution it is necessary to consider a set of proteins with quantitative measurements of their functional efficiency. However, it is usually difficult to precisely characterize protein molecular function and especially to quantify its optimality across different proteins. Fortunately, based on the Enzyme Commission (EC) four-digit classification scheme (Webb 1992), diverse enzymatic functions have been well defined. Moreover, catalytic rates of many enzymes have been measured using accurate low throughput biochemical experiments (Chen and Vitkup 2007; Wittig et al. 2018; Chang et al. 2021). Two biochemical parameters, k_{cat} and k_{cat}/K_M, that characterize catalytic activities can be used to evaluate the functional efficiency of enzymes. The first-order kinetic rate constant, k_{cat}, quantifies the speed (turnover) of enzymatic reactions at saturating concentrations of substrates, and the specificity constant, k_{cat}/K_M, quantifies the second-order reaction rate at ligand concentrations substantially lower than the Michaelis constant, K_M. Enzymes achieve their amazing catalytic efficiency primarily by stabilizing transition states of corresponding chemical reactions (Abeles et al. 1992). However, depending on the chemical properties of substrates and the nature of catalyzed biochemical interconversions, it is much easier to achieve high kinetic rates for some enzymatic classes than for others. This makes the comparison of absolute kinetic rates between different enzymatic classes not very informative. Therefore, following previous studies (Davidi et al. 2018), we quantified the enzymatic functional optimality using relative catalytic rates. Specifically, we normalized absolute kinetic constants by the highest catalytic constants experimentally measured for enzymes from the same reaction class, i.e. enzymes sharing all four digits of the EC classification. The catalytic rates normalized in this way reflect the extent to which the rate constants

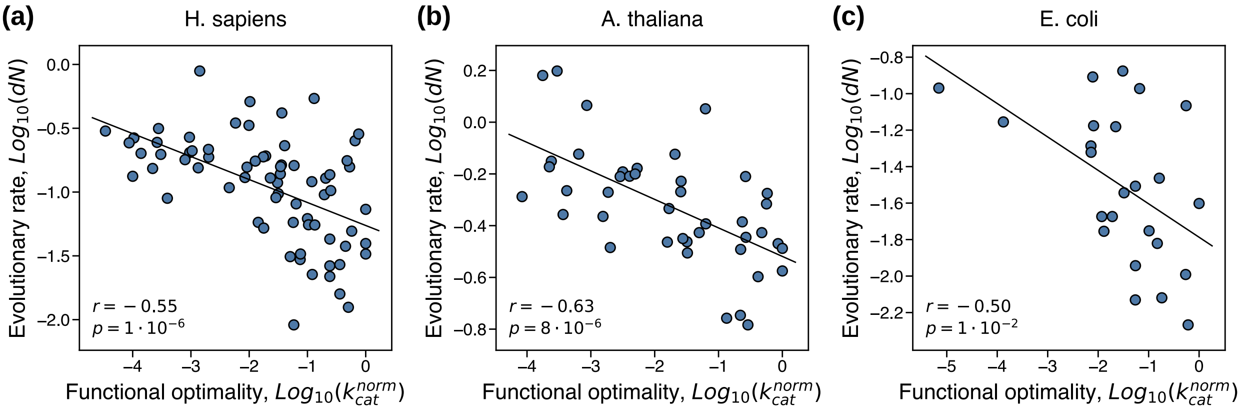


Fig. 1. The correlation between protein functional optimality and the rate of protein evolution. Each point in the figures represents an enzyme from a) *H. sapiens* ($n = 70$), b) *A. thaliana* ($n = 42$), and c) *E. coli* ($n = 24$). Protein functional optimality was estimated using the normalized kinetic constant, k_{cat}^{norm}, which quantifies the turnover catalytic rate relative to the maximal known rate for the same reaction class. Evolutionary rate, *dN*, was calculated as the number of non-synonymous substitutions accumulated during the divergence of closely related orthologs per non-synonymous site. Spearman's correlation coefficients and *P*-values are shown in each figure.

deviate from the maximal known rates from the same enzymatic class. To accurately estimate enzymatic efficiencies using the normalized rate constants we only considered EC classes with a certain minimal number of different enzymes for which kinetic constants were experimentally measured (**Methods**). As we describe below, analyses based on the normalized kinetic rate constants (k_{cat}^{norm} and $(k_{cat}/K_M)^{norm}$) reveal insightful correlations between enzymatic optimality and protein evolutionary rates.

To analyze enzymatic functional optimality, we used a large collection of experimental k_{cat} and k_{cat}/K_M measurements available in the Brenda (Chang et al. 2021) and Sabio-RK databases (Wittig et al. 2018). The largest number of kinetic constants in these databases were measured for enzymes from *H. sapiens*, *A. thaliana*, and *E. coli*. Available experimental measurements of kinetic constants from other organisms allowed us to estimate functional optimality for a substantial number of enzymes from these three species. Notably, in all species we observed substantial negative correlations between the rate of protein evolution, quantified as the rate of non-synonymous substitutions between orthologs in closely related organisms, *dN*, and functional optimality, quantified using either k_{cat}^{norm} (Spearman's $r = -0.55$, $p = 1 \cdot 10^{-6}$, for *H. sapiens*; $r = -0.63$, $p = 8 \cdot 10^{-6}$, for *A. thaliana*; $r = -0.50$, $p = 1 \cdot 10^{-2}$, for *E. coli*; Fig. 1) or $(k_{cat}/K_M)^{norm}$ (supplementary fig. S1, Supplementary Material online). By analogy to the ER (expression-evolutionary rate) correlation, we refer to this correlation as KR, i.e. the correlation between the normalized kinetic rate (K) and the rate of protein evolution (R). We also evaluated the strength of selection against enzyme mutations using the ratio of non-synonymous to synonymous substitution rates, dN/dS (Li et al. 1985), and found that all enzymes considered in the analysis evolve under purifying selection, i.e. with $dN/dS < 1$. In all species, dN/dS was also significantly and negatively correlated with the level of functional optimality (Spearman's $r = -0.46$, $p = 6 \cdot 10^{-5}$, for *H. sapiens*; $r = -0.44$, $p = 4 \cdot 10^{-3}$, for *A. thaliana*; $r = -0.46$, $p = 2 \cdot 10^{-2}$, for *E. coli*; supplementary fig. S2, Supplementary Material online). The KR and K-dN/dS correlations were observed across a wide range (~5 orders of magnitude) of protein optimality levels, and demonstrate that higher optimality of protein molecular function indeed usually leads to substantially slower rates of protein evolution. Slower evolutionary rates of proteins with highly optimized molecular function likely result from the additional constraints required to maintain protein sequence, three-dimensional structure, and protein dynamics necessary for efficient function and catalysis (Konate et al. 2019). We note that the explanatory power of protein functional optimality for predicting evolutionary rates is substantially higher compared to multiple other protein biochemical and biophysical properties, such as stability, solubility, and stickiness, which usually account for only a small percentage (1% to 5%) of the evolutionary rate variance (Plata et al. 2010; Plata and Vitkup 2018; Usmanova et al. 2021).

The evolutionary pressure to optimize protein function should be especially strong for highly expressed proteins, as such optimization allows cells to substantially reduce the number of expressed proteins. The FORCE model also suggests that in multicellular organisms both the pressure for functional optimization and the ER correlation should be stronger in the tissues with high protein production costs. Several possible scenarios can make certain cells and tissues particularly sensitive to protein expression costs. One such scenario involves fast-growing cells where protein expression is likely to be a major burden. Constitutive protein expression, maintained due to constant protein turnover, is itself a major source of cellular energy consumption (Buttgereit and Brand 1995; Rolfe and Brown 1997). Therefore, another potential scenario of cells with high expression costs is either growing or non-

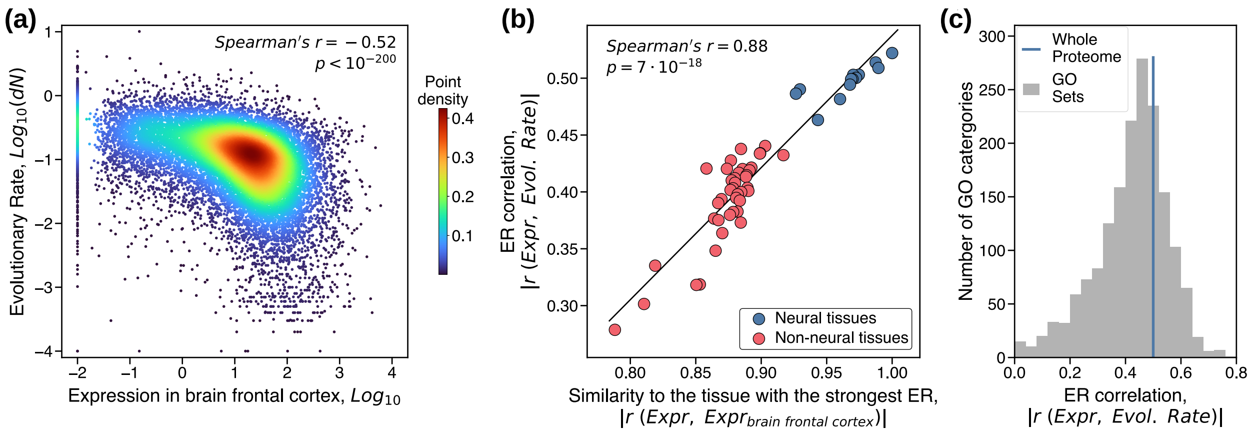


Fig. 2. The relationship between expression in human tissues and the rate of protein evolution. a) The correlation between gene expression in the brain frontal cortex and evolutionary rates of the corresponding human proteins. Evolutionary rates, *dN*, were calculated as the number of non-synonymous substitutions accumulated during the divergence of closely related orthologs per non-synonymous site (see **Methods**). Each point in the plot represents a human protein ($n = 18{,}619$), and the colors represent the point density. b) The correlation between ER values across human tissues and the similarity of tissues' genes expression to the frontal cortex; similar results were obtained for other animals (supplementary fig. S3, d to f, Supplementary Material online). The similarity between tissues' expression profiles was quantified using the Spearman's correlation. Blue points ($n = 13$) represent human neural tissues, and red points ($n = 40$) represent non-neural tissues; the linear regression and the Spearman's correlation coefficient were calculated based on all 53 tissues. c) The distribution of the ER correlation strength, based on the brain frontal cortex expression, across different Gene Ontology (GO) categories each containing at least 100 human genes; blue vertical line indicates the ER correlation of the whole proteome.

growing cells with substantial and persistent energy expenditures. To explore these scenarios, we next investigated which tissues and cellular processes are typically associated with stronger ER correlations and with more pronounced protein functional optimization in multicellular organisms.

The Rate of Protein Evolution and Gene Expression in Animals

Previous studies in animals demonstrated (Duret and Mouchiroud 2000; Zhang and Li 2004; Wang et al. 2007) that proteins highly expressed in the brain generally evolve slowly (Fig. 2a, supplementary fig. S3, a to c, Supplementary Material online). Thus, we first investigated to what extent expression in non-neural tissues further constrains protein evolutionary rates. We selected for this analysis several model organisms: *Homo sapiens* (Mele et al. 2015), *Mus musculus* (Söllner et al. 2017), *Drosophila melanogaster* (Leader et al. 2018), and *Caenorhabditis elegans* (Spencer et al. 2011); these organisms span ~800 million years of divergence time, and their mRNA expression data are available across diverse tissues and cell types. Application of multivariable regression analysis showed that only a small additional fraction (<4%) of the protein evolutionary rate variance can be explained by considering expression in all tissues compared to expression in neural tissues only (supplementary table S1, Supplementary Material online, **Methods**). We observed that the tissues with the weakest ER in humans (the testis, blood, and liver) tend to have expression profiles most distant from those in the brain (supplementary table S2, Supplementary Material online). Furthermore, we found that the strength of tissue-specific ER correlations in all species could be largely explained by the similarity between gene expression in a particular tissue and in the neural tissue with the strongest ER (Fig. 2b, supplementary fig. S3, d to f, Supplementary Material online), confirming that the expression-based evolutionary constraints are primarily dominated by neural tissues. We next explored the gene expression breadth across tissues (Duret and Mouchiroud 2000; Park and Choi 2010), as this characteristic of global gene expression was suggested to be another important factor in constraining evolutionary rates in multicellular species (see **Methods**). We found that the expression breadth correlates with evolutionary rates stronger than gene expression in many non-neural tissues, but substantially weaker than expression in neural tissues (supplementary table S1, Supplementary Material online). Notably, the expression breadth explained little additional variance of evolutionary rates (~1% for all species) when combined in the regression analysis with neural expression.

It has been previously demonstrated that protein expression significantly correlates with the rate of protein polymorphisms in bacteria (Feugeas et al. 2016). Similarly, based on the analysis of human polymorphism data (The 1000 Genomes Project Consortium 2015), we observed that expression in neural tissues strongly correlates not only with the rate of interspecies protein evolution, but also with the per-site frequency of polymorphisms across proteins in the human population (supplementary fig. S4, Supplementary Material online). This suggests that expression in neural tissues plays a dominant role in constraining protein evolution both between and within populations.

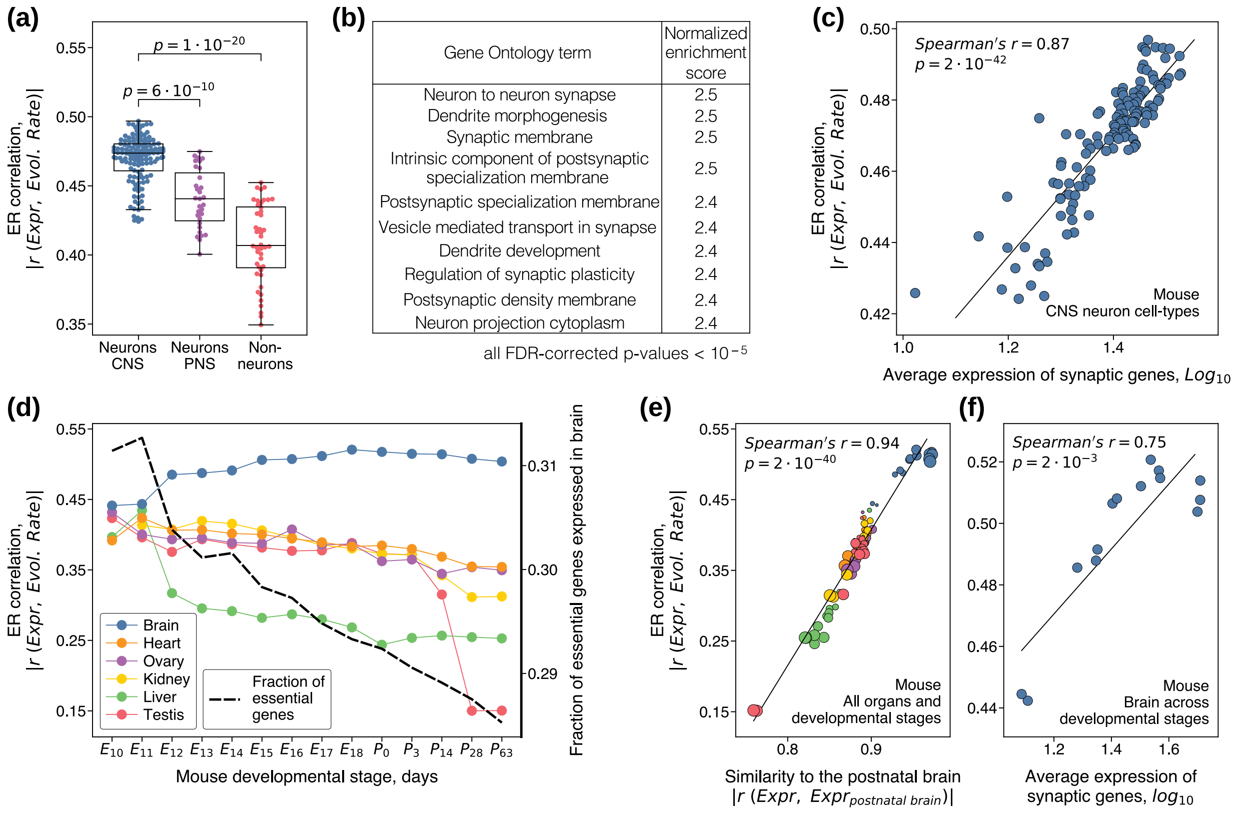
We investigated next the ER correlation for groups of genes associated with specific biological and cellular functions. To that end, for each of the ∼1,600 Gene Ontology (GO) categories (including molecular functions, biological processes, and cellular components) with at least 100 human genes, we calculated the ER correlation using only genes annotated with each of these GO terms. The ER correlation, calculated based on gene expression in the brain frontal cortex, was significant for 97% of these function-specific gene sets (Fig. 2c). As the expression variance within function-specific gene sets was typically smaller compared to the whole proteome, ER within individual GO categories (the median ER strength of 0.44) was also slightly smaller than for the entire proteome (the ER strength of 0.52). Only a very small fraction of GO categories showed non-significant ER correlations. For example, almost all genes belonging to the GO term "Oxidative Phosphorylation" are highly expressed and have relatively low evolutionary rates, while genes belonging to the GO term "Olfactory Receptor Activity" have low expression and high evolutionary rates; as a result, both of these GO terms have non-significant ER (supplementary table S3, Supplementary Material online). Despite diverse functions represented by different GO categories, the ranking of tissues by the ER strength was mostly preserved across them, with brain tissues demonstrating the strongest ER for 82% of the individual GO categories. Overall, this analysis of GO annotations demonstrates that the ER correlation is a general property of almost all protein molecular and cellular functions.

To further explore why the strongest ER correlation is observed in neural tissues, we first established that this effect is not primarily mediated by evolutionary properties of neural- or brain-specific genes. The removal of these genes from the analysis did not substantially weaken the ER correlation in neural tissues; for example, excluding 10% of the most neural-specific genes changed the ER correlation of the remaining genes by less than 3% in all species (**Methods**). Moreover, even for the subsets of genes most specific to non-neural tissues, the ER correlations calculated based on their expression in the brain were almost always substantially stronger than based on expression in non-brain tissues (supplementary figs. S5 and S6, Supplementary Material online; **Methods**). Strong ER correlations are also unlikely to result from the difference in the number of expressed genes in various tissues. Setting the abundance of lowly-expressed genes to zero, in order to equalize the number of expressed genes across tissues, decreased the ER correlations by ∼1% in the neural tissues and also preserved the ranking of tissues by ER (Spearman's $r > 0.98$, $p < 4 \cdot 10^{-12}$ for all species; **Methods**). Because it was previously demonstrated that the fraction of essential genes is similar across mouse tissues (Cardoso-Moreira et al. 2019), the observed differences in the ER correlation are also unlikely to originate from the higher essentiality of brain-expressed genes. Therefore, the strong correlation between evolutionary rate and gene expression in neural tissues is likely to be mediated not by expression of brain-specific genes themselves but by some inherent functional properties of neural cells that affect evolution of all proteins expressed in these cells.

The Role of Neuronal Gene Expression in Constraining Protein Evolution

To investigate the cellular properties that may underlie the ER correlations we took advantage of cell-type-specific brain expression data (Davie et al. 2018; Saunders et al. 2018; Zeisel et al. 2018; Sugino et al. 2019). We first analyzed the single-cell transcriptome dataset by Zeisel et al. (2018); the dataset covers hundreds of cell types including neurons and non-neurons across the entire mouse brain (**Methods**). Consistent with previous analyses (Hu et al. 2020), we found that neurons generally have significantly stronger ER correlations than non-neuron brain cells (Mann–Whitney test $p = 1 \cdot 10^{-20}$; Fig. 3a), with the ER correlations in central nervous system (CNS) neurons significantly stronger than in neurons of the peripheral nervous system (PNS) (Mann–Whitney test $p = 6 \cdot 10^{-10}$). Notably, the strength of the ER correlation varied between different types of neurons, and we leveraged this variability to explore which specific cellular functions are usually upregulated in the neuron types associated with stronger ER correlations. To that end, we ranked all mouse genes based on how strongly their individual expression correlates with the ER strength across all CNS neuron types. We then used the gene set enrichment analysis (GSEA) (Subramanian et al. 2005) to identify functional GO categories that were associated with higher gene rankings (see **Methods**, supplementary table S4, Supplementary Material online). The GSEA analysis showed that neurons with stronger ER correlations tend to have higher expression of genes associated with synaptic functions and related cellular processes (Fig. 3b). An alternative GSEA approach, which ranked mouse genes based on the differential expression between CNS neuron types with high and low ER strengths, also implicated similar GO categories (**Methods**, supplementary table S4, Supplementary Material online). Consistent with the GSEA results, we observed a strong correlation between the average expression of synaptic genes and the strength of ER correlation across neuron types (Spearman's $r = 0.87$, $p = 2 \cdot 10^{-42}$; Fig. 3c). We again note that the correlation patterns identified by the GSEA analysis were not due to the synaptic genes themselves, as removing synapse-associated genes or genes from all upregulated GO categories from the analysis, i.e. using these genes only in ranking but not in the ER calculations across neuron types, did not substantially change the GSEA results (**Methods**, supplementary table S4, Supplementary Material online). Instead, it is likely that innate cellular properties of neurons with strongly upregulated synaptic functions make evolutionary rates of all proteins especially sensitive to expression levels in these cell types.

Next, we confirmed the correlation between the expression of synaptic genes and the ER strength using several independent datasets. First, we analyzed two additional



Gene Ontology term	Normalized enrichment score
Neuron to neuron synapse	2.5
Dendrite morphogenesis	2.5
Synaptic membrane	2.5
Intrinsic component of postsynaptic specialization membrane	2.5
Postsynaptic specialization membrane	2.4
Vesicle mediated transport in synapse	2.4
Dendrite development	2.4
Regulation of synaptic plasticity	2.4
Postsynaptic density membrane	2.4
Neuron projection cytoplasm	2.4

Fig. 3. Functional properties of mouse brain cells and developmental stages that are associated with stronger ER correlations. a) The strength of ER correlations across different cell types in the mouse nervous system ($n = 207$). CNS neurons are shown in blue, PNS neurons in purple, and non-neuronal brain cells in red. The box plots show the median, the upper and lower quartiles of the ER strength, and the whiskers show the minimum and maximum values excluding outliers; *P*-values were calculated using the Mann–Whitney *U* test. b) The GO terms associated with stronger ER in mouse CNS neurons. The top 10 GO terms ranked by the strength of the normalized enrichment score (see **Methods**) are shown; False Discovery Rate (FDR)-corrected *P*-values < 10^{-5} for all presented GO terms. The complete list of significantly associated GO terms is provided in the supplementary table S4, Supplementary Material online. c) The correlation between the average expression of genes from the Synapse GO term (GO:0045202) and the ER strength across CNS cell types; each point in the figure represents a mouse CNS neuron type ($n = 132$). d) The strength of ER correlation for mouse tissues across prenatal and postnatal developmental stages. Color lines represent different organs, the *x*-axis shows mouse developmental stages, and the left *y*-axis shows the ER correlation strength. The right *y*-axis shows the fraction of essential genes (represented by the dashed black line) significantly expressed in the brain at different developmental stages. e) The correlation between the similarity of samples' gene expression profiles with the postnatal brain expression (*x*-axis) and the ER strength across mouse organs and developmental stages (*y*-axis); the similarity between expression profiles was quantified using Spearman's correlation. Postnatal brain expression was calculated as the average over all corresponding postnatal samples. The color of each dot ($n = 84$) represents an organ according to the legend in (d), and the size of each dot represents the developmental stage, with smaller dots corresponding to earlier developmental stages. f) The correlation between the average expression (*x*-axis) of genes from the Synapse GO term (GO:0045202) and the ER strength (*y*-axis); each point in the figure represents a different mouse brain developmental stage ($n = 14$).

comprehensive cell-type-specific transcriptomes of the mouse brain. One transcriptome was obtained using single-cell sequencing (Saunders et al. 2018), while the other was obtained using the bulk RNA sequencing of cell populations distinguished by their genetic and anatomical markers (Sugino et al. 2019). The GSEA analysis applied to these datasets confirmed the significant association between the upregulation of synapse-related functions and the strength of ER (supplementary fig. S7, Supplementary Material online, supplementary table S4, Supplementary Material online). We also analyzed a region-specific gene expression dataset that covers the entire mouse brain (∼100 brain regions) and which was obtained using in situ hybridization (Lein et al. 2007). This dataset allowed us to correlate the ER strength with brain regions' physiological properties that were measured in the same three-dimensional coordinate system describing the mouse brain (Erö et al. 2018; Murakami et al. 2018; Zhu et al. 2018). Interestingly, we found that the ER strength was strongly correlated with the density of synapses (Zhu et al. 2018) across the mouse brain (Spearman's $r = 0.66$, $p = 2 \cdot 10^{-13}$), but not with the density of neurons (Erö et al. 2018; Spearman's $r = -0.051$, $p = 0.6$) or the overall cellular density (Murakami et al. 2018; Spearman's $r = -0.15$, $p = 0.14$) (supplementary fig. S8, Supplementary Material online). Finally, we analyzed a single-cell expression dataset from the *D. melanogaster* brain (Davie et al. 2018), and observed that in the non-vertebrate species protein evolutionary rates also correlate significantly stronger with expression in neurons than in other brain cells (Mann–Whitney test $p = 9 \cdot 10^{-4}$; supplementary fig. S7g, Supplementary Material online). The GSEA analysis applied to the *D. melanogaster* expression dataset also showed a significant association between the ER strength and the upregulation of synaptic and neuropeptide signaling functions, indicating the generality of these patterns in diverse species (supplementary fig. S7, h and i, Supplementary Material online; supplementary table S4, Supplementary Material online).

Functional properties of cells and their gene expression profiles vary not only between adult tissues and cell types but also across developmental stages, with especially rapid changes observed during embryogenesis. To investigate how the strength of ER changes during development, we analyzed a temporal expression dataset that covers multiple mouse organs and includes both prenatal and postnatal developmental periods (Cardoso-Moreira et al. 2019). In agreement with previous observations (Hu et al. 2020), we found that expression in neural tissue becomes more strongly correlated with protein evolutionary rates as embryonic development progresses (Fig. 3d). The maximal ER in the brain was reached around birth, without substantial further changes during the postnatal developmental stages. Qualitatively different behaviors were observed for non-neural tissues, for which the ER correlation monotonically decreased from the early to late developmental stages (Fig. 3d). This pattern likely arises because tissues' expression profiles are more similar during the early developmental stages, but continuously diverge as the development progresses (Cardoso-Moreira et al. 2019). As in the analysis of adult animal tissues (Fig. 2b, supplementary fig. S3, d to f, Supplementary Material online), we found that the ER strength calculated for samples from different organs and different developmental stages is highly correlated with the similarity of samples' gene expression to expression in the adult brain (Fig. 3e).

We next investigated the ER variability across brain developmental stages. We performed the GSEA analysis, which showed that the GO terms associated with stronger ER were primarily the functional categories also associated with stronger ER in adult neurons (supplementary table S4, Supplementary Material online). Specifically, we found that the ER strength strongly correlates with synaptic gene expression across the developmental stages (Spearman's $r = 0.75$, $p = 2 \cdot 10^{-3}$; Fig. 3f). As we stated previously, the strong ER correlation in the developed brain is unlikely to be mediated by gene essentiality. Essential genes (Koscielny et al. 2014), on average, evolve twice slower than non-essential genes (Rocha and Danchin 2004), but demonstrate substantially weaker ER correlations (supplementary fig. S9, Supplementary Material online). Interestingly, similar to the ER of non-essential genes, the ER of essential genes in the brain is weaker during early development (supplementary fig. S9, Supplementary Material online), despite the fact that essential genes are more highly expressed at those developmental stages (Cardoso-Moreira et al. 2019) (Fig. 3d, dashed black line, **Methods**).

Overall, these results demonstrate that protein evolutionary rates in animals correlate more strongly with gene expression in developed neurons, especially in neurons with upregulated molecular and cellular functions related to synaptic activities. Our analysis also suggests that this effect is not primarily due to synaptic genes themselves, but that it is likely mediated by the functional properties of neurons in which synapse-related genes are highly expressed. Due to the generality of these results in animals, it is interesting to investigate cellular processes and functions that primarily affect protein evolutionary rates in multicellular organisms without neural tissues. Thus, we next considered the tissue-specific ER correlations and associated cellular processes in plants.

The Rate of Protein Evolution and Gene Expression in Plants

We investigated the ER correlations in plants using multitissue RNA-seq data from three angiosperm species: *Zea mays* (corn) (Stelpflug et al. 2016), *Arabidopsis thaliana* (Klepikova et al. 2016), and *Glycine max* (soybean) (Shen et al. 2014). As was reported previously, the similarity of plant tissues' transcriptomes often reflects not only the relatedness of their morphological origins but also the similarity of their developmental stages (Klepikova et al. 2016; Stelpflug et al. 2016). We confirmed this observation in the considered plant species based on hierarchical clustering (**Methods**) of the tissues' expression data (Fig. 4). This analysis resulted in distinct expression clusters representing samples from roots, leaves, stems, seeds, flowers, meristems, and also clusters that included multiple young or growing tissues of diverse origin. For example, the uppermost cluster in the dendrogram for corn (top in Fig. 4a) combines the samples from seedlings, root axes, leaf buds, developing seeds or flowers (see supplementary table S2, Supplementary Material online for samples to cluster assignments). In contrast to animals, in plants we did not observe morphologically similar tissue types that have universally strong ER correlations (Fig. 4). However, plants' samples with stronger ER tended to include young and growing tissues, while senescent tissues always exhibited relatively low ER correlations. Similar to animals, the observed patterns were not primarily due to the genes specific to growing plant tissues, as removal of such genes did not substantially affect the strength of ER (**Methods**); for example, removing 10% of the genes most specific to the growing tissues changed the ER correlations of the remaining genes in the plant tissues with strong ER by less than 3%. This analysis again suggests that the functional properties of fast-growing plant tissues likely make protein evolutionary rates especially sensitive to expression in the corresponding cells and organs.

To understand the functional properties of plant cells associated with strong ER, we again used the GSEA enrichment analysis. In all three considered plant species we found that similar GO categories are usually associated

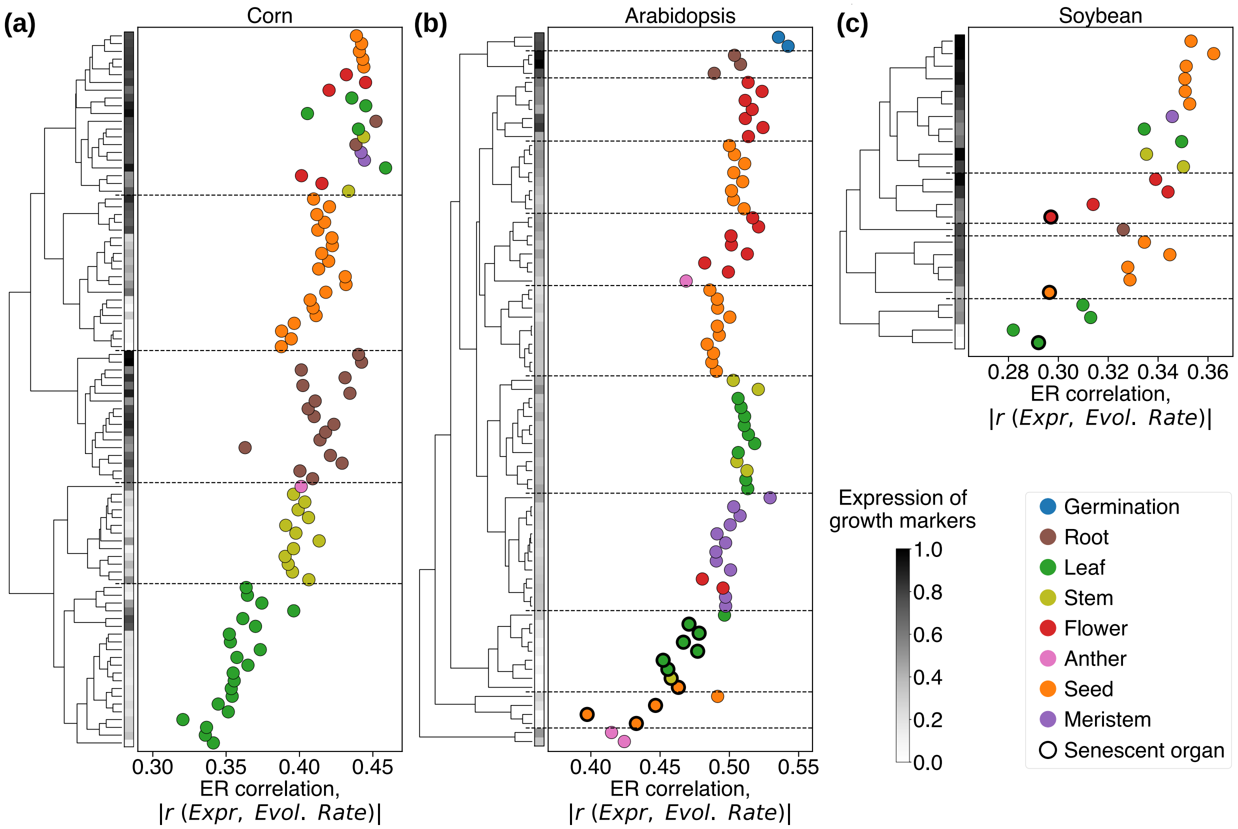


Fig. 4. The relationship between gene expression in plant tissues and the rate of protein evolution. The data are shown for transcriptomes of a) corn (Stelpflug et al. 2016), b) Arabidopsis (Klepikova et al. 2016), and c) soybean (Shen et al. 2014). Each point in the figure represents a plant tissue sample, point colors represent different plant tissue types described in the legend; senescent samples are shown as black-edged circles. The *x*-axis represents the strength of the ER correlation. The left panel of each plot shows the hierarchical clustering dendrogram of the plant transcriptomes, with the clustering distance metric calculated as one minus the squared Pearson's correlation coefficient between samples' expression profiles. The horizontal dashed lines separate major clusters of the dendrogram. The vertical gray scale colormaps on the left side of the plots show the scaled average expression of cell growth markers in the corresponding plant tissues. The genes strongly upregulated in fast-growing root cells (Huang and Schiefelbein 2015) were used as the growth markers for Arabidopsis (see **Methods**); the orthologs of the Arabidopsis growth markers were used as the growth markers for corn and soybean.

with stronger ER correlations (supplementary fig. S10, Supplementary Material online; supplementary table S4, Supplementary Material online). These upregulated GO terms primarily represent growth-related functional categories, such as translation, cell wall biosynthesis, and microtubule cytoskeleton organization/movement (Fig. 5a). The implicated functional categories suggest that strong ER correlations in plants are usually associated with cellular elongation and growth. The growth of plant organs is known to be initiated in the zone of undifferentiated meristematic cells and consists of three consecutive stages: cells division, elongation and differentiation (Taiz et al. 2015). Therefore, we used the distinct gene markers of these growth stages (Huang and Schiefelbein 2015) to further investigate how the marker's average gene expression correlates with ER across tissues (**Methods**). This analysis demonstrated that expression of the cellular elongation markers strongly correlates with the ER strength in all three plant species (Spearman's $r = 0.72$, $p = 8 \cdot 10^{-16}$, for corn; $r = 0.75$, $p = 3 \cdot 10^{-15}$, for Arabidopsis; $r = 0.83$, $p = 4 \cdot 10^{-7}$, for soybean; Fig. 5, b to d). The expression of the cell division markers showed a weaker and less significant correlation with the ER strength (Spearman's $r = 0.48$, $p = 1 \cdot 10^{-6}$, for corn; $r = 0.15$, $p = 0.2$, for Arabidopsis; $r = 0.36$, $p = 0.08$, for soybean; supplementary fig. S11, Supplementary Material online). Finally, the correlation between the cell differentiation markers and ER was either not significant or was significant in the direction opposite to the other two markers (Spearman's $r = -0.63$, $p = 2 \cdot 10^{-11}$, for corn; $r = -0.02$, $p = 0.9$, for Arabidopsis; $r = -0.32$, $p = 0.1$ for soybean), confirming a substantial decrease of the ER strength for plant tissues entering the differentiation stage.

While different cellular functions are associated with strong ER correlations in animals and plants, our results suggest that the strongest correlations are usually observed in tissues with high expression costs. In plants, our analysis implicates cells from various tissues that are rapidly growing, and therefore likely prioritizing their carbon and energy resources for novel protein production. In animals, brain synaptic activity requires a substantial and persistent energy

(a)

GO Term	Normalized enrichment score		
	Corn	Arabidopsis	Soybean
Structural constituent of ribosome	2.9	3.9	2.6
Translation	1.8	3.6	2.1
Microtubule cytoskeleton organization	2.9	2.8	2.9
Microtubule-based movement	2.6	2.9	2.8
Kinesin complex	2.6	2.7	2.8
Mitotic cell cycle	2.8	2.5	2.7
Spindle assembly	2.4	2.5	2.9
Plant-type cell wall	1.7	2.6	1.8
Cellulose biosynthetic process	1.7	2.4	2.4
Intra-Golgi vesicle-mediated transport	2.8	1.7	2.5

all FDR-corrected p-values < 0.05

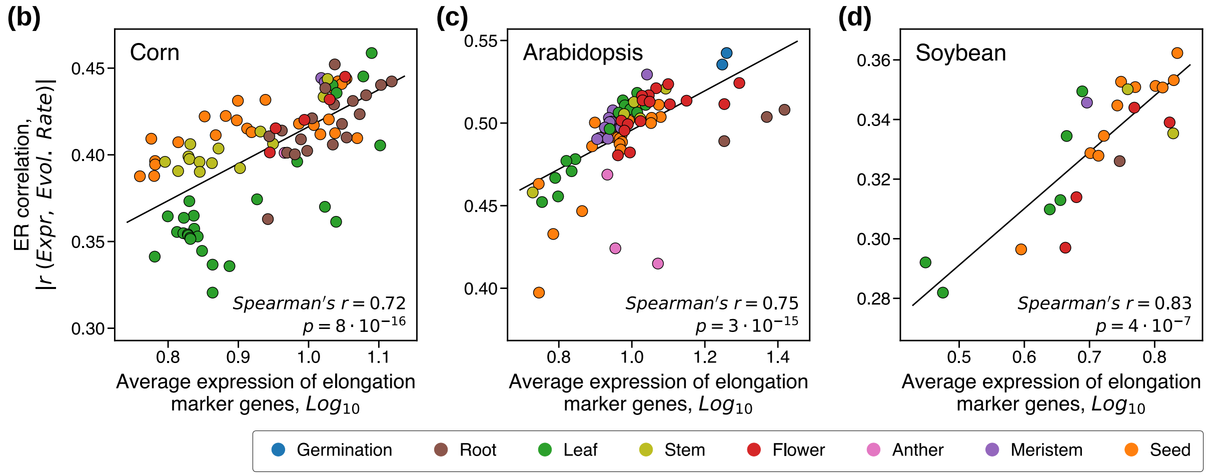


Fig. 5. Functional properties of plant tissues that are associated with stronger ER correlations. a) GO categories significantly associated with stronger ER correlations in the three considered plant species: corn, Arabidopsis, and soybean. Normalized enrichment scores for the top 10 representative GO terms are shown; FDR-corrected *P*-values <0.05 for all presented GO terms. The complete list of upregulated GO categories is provided in supplementary table S4, Supplementary Material online. b to d) The relationship between the average expression of the elongation growth markers and the tissue-specific ER strength for b) corn ($n = 92$), c) Arabidopsis ($n = 79$), and d) soybean ($n = 25$). Each point in the figure represents a plant tissue, and point colors represent the tissue types described in the legend.

supply (Harris et al. 2012). Thus, it may be particularly important to reduce the protein expression burden for neurons with high densities of synaptic connections. Having identified the tissues with the strongest ER correlations, we next investigated to what extent the correlation between protein functional optimality and evolutionary rate may mediate the ER correlation and its variability across tissues.

The Role of Protein Functional Optimization in Mediating the ER Correlation

To investigate the relationship between the optimization of protein function and the ER correlation, we considered next the sets of *H. sapiens*, *A. thaliana*, and *E. coli* enzymes with estimated levels of their functional optimality (Fig. 1). First, we confirmed that for the enzymes from these sets the ER correlation is significant and similar in strength to the ER correlation for the entire proteomes (Spearman's $r = -0.64$, $p = 3 \cdot 10^{-9}$, for *H. sapiens*; $r = -0.61$, $p = 2 \cdot 10^{-5}$, for *A. thaliana*; $r = -0.75$, $p = 3 \cdot 10^{-5}$, for *E. coli*, supplementary fig. S12, Supplementary Material online). This result suggests that the mechanisms underlying the whole-proteome ER correlation also play a similar role in the evolution of these specific sets of proteins. The FORCE mechanism of the ER correlation proposes that highly expressed proteins are more functionally optimized to relieve the burden associated with their production costs. Consistent with this model, we observed in all three species significant correlations between protein expression level and protein functional optimality (Spearman's $r = 0.52$, $p = 4 \cdot 10^{-6}$, for *H. sapiens*; $r = 0.45$, $p = 3 \cdot 10^{-3}$, for *A. thaliana*; $r = 0.46$, $p = 2 \cdot 10^{-2}$, for *E. coli*; Fig. 6, a to c). By analogy to the ER and KR correlations described above, we refer to this revealing correlation as EK, i.e. the correlation between expression (E) and kinetic constants (K).

Our analyses of protein expression across tissues demonstrated that the rate of protein evolution is especially sensitive to expression in several specific cell types and tissues, such as neurons in animals (Fig. 2) and actively growing tissues in plants (Fig. 4). The tissues most sensitive to expression costs are likely to exert the highest selective pressure to optimize protein function. As a result, both the EK and ER correlations should be stronger in tissues with high expression costs and weaker in other tissues. In agreement with this prediction, in both animals and plants we observed significant correlations between the strengths of ER and EK calculated across tissues (Spearman's $r = 0.60$, $p = 2 \cdot 10^{-6}$, for

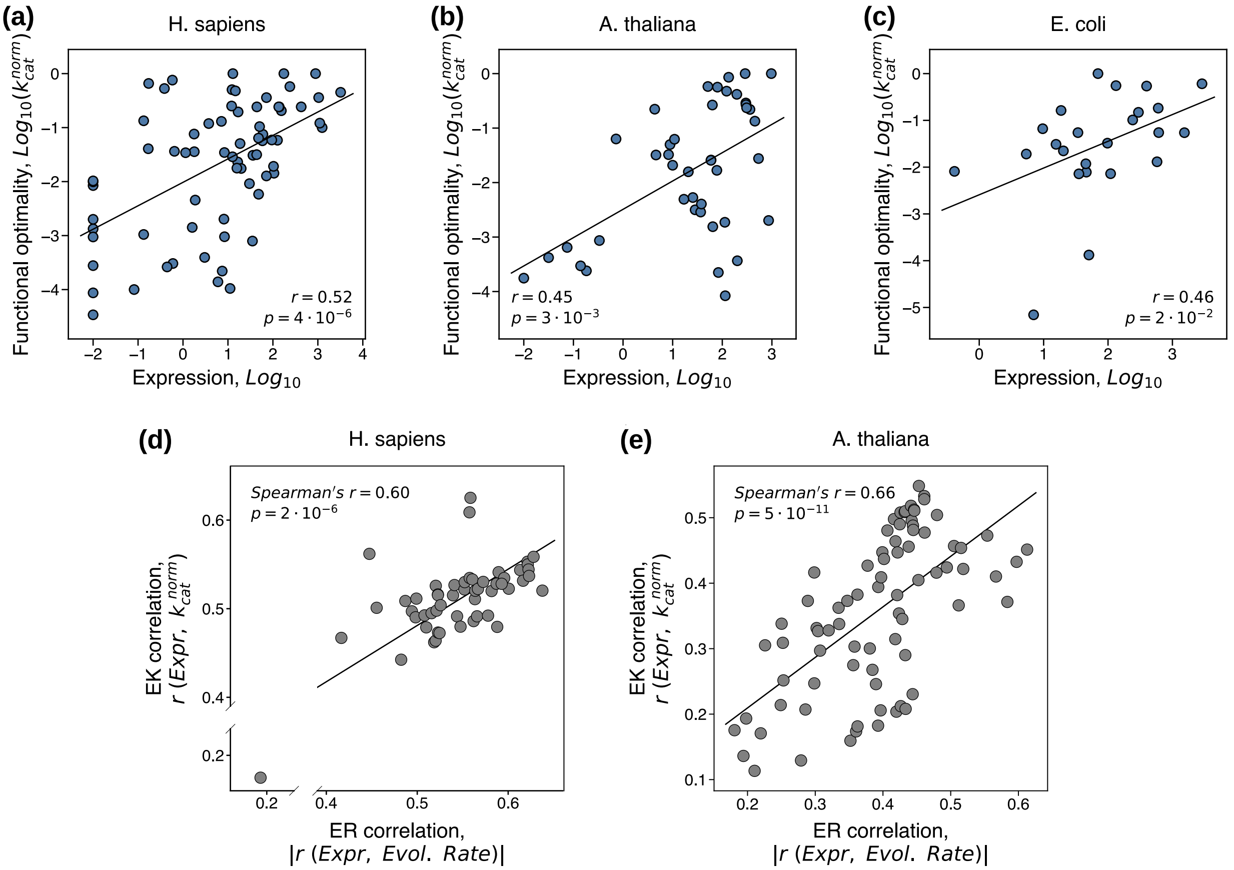


Fig. 6. The relationship between the level of protein functional optimality, expression, and the rate of protein evolution. a to c) The correlation between protein expression level and protein functional optimality (the EK correlation). Each point on the plots represents an enzyme from a) *H. sapiens* ($n = 70$), b) *A. thaliana* ($n = 42$), and c) *E. coli* ($n = 24$). Protein functional optimality was estimated using the normalized kinetic constant, k_{cat}^{norm}, which quantifies the turnover catalytic rate relative to the maximal known rate for the same reaction class. For multicellular species, the expression in the tissue with the strongest ER for enzymes was used (the brain basal ganglia for *H. sapiens* and the seedling root for *A. thaliana*). (d, e) The relationship between the ER correlation (the correlation between expression and evolutionary rate) and the EK correlation (the correlation between expression and protein functional optimality, k_{cat}^{norm}). Each point represents a tissue from d) *H. sapiens* (53 tissues) or e) *A. thaliana* (79 tissues); the strength of ER and EK correlations were quantified across tissues using Spearman's correlation.

H. sapiens; $r = 0.66, p = 5 \cdot 10^{-11}$, for *A. thaliana*; Fig. 6, d and e), with the highest correlations observed in the brain for *H. sapiens* and in fast-growing tissues for *A. thaliana*.

Finally, we quantitatively investigated what fraction of the ER correlation can be explained by the variability in protein functional optimality. To that end, we used the semi-partial correlations to calculate the unique and shared contributions of two independent variables, expression and protein functional optimality, in explaining the variance of protein evolutionary rates. This analysis demonstrated that after accounting for protein functional optimality, quantified using k_{cat}^{norm}, the fraction of the evolutionary rate variance explained by expression substantially decreases: by ~1/2 for *H. sapiens* (from 41% to 17%), ~2/3 for *A. thaliana* (from 38% to 14%), and ~1/3 for *E. coli* (from 56% to 34%); similar results were obtained for the multicellular species when controlling for $(k_{cat}/K_M)^{norm}$ (supplementary fig. S13, a to c, Supplementary Material online). The observed decreases in the variance explained are due to the shared fractional contribution of protein expression and functional optimality to the variance of evolutionary rate, and the remaining fractions represent their unique contributions (Fig. 7). The slowing of evolutionary rates of highly expressed proteins is likely to be mediated by stronger selection to maintain functionally optimal protein sequences. Indeed, in all three species, about half of the correlation between expression and dN/dS can be also attributed to the functional efficiency as an intermediate variable (supplementary fig. S13, d to i, Supplementary Material online). The large explanatory effect sizes of protein functional optimality in explaining the evolutionary rate variability are especially remarkable because our k_{cat}^{norm}-based optimality estimations relied on a simple heuristic normalization procedure. We note that the fractions of ER unexplained by k_{cat}^{norm} do not necessarily indicate that they are unrelated to protein function, because kinetic constants, such as k_{cat}, are clearly not the only parameters optimized in protein evolution. The refinements of multiple other functional properties, such as the efficiency of allosteric regulation, covalent modification, and protein–protein binding,

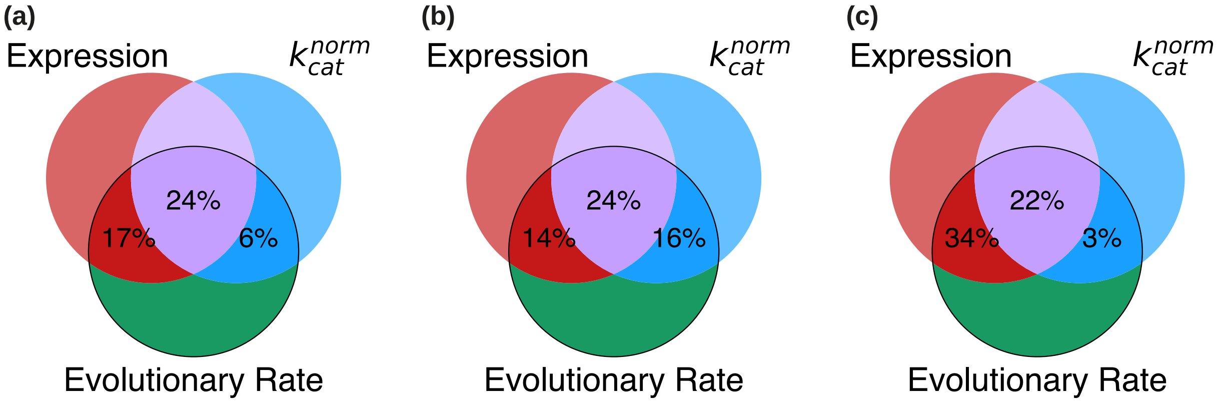


Fig. 7. The fractions of the evolutionary rate variance explained by protein functional optimality and expression. The Venn diagrams show for a) *H. sapiens*, b) *A. thaliana*, and c) *E. coli*, the fractions of the evolutionary rate variance explained by expression and protein functional optimality; functional optimality was quantified using the normalized kinetic constant k_{cat}^{norm}. The unique and shared contributions were estimated using semi-partial correlations (see **Methods**). The two-way intersections (dark red and dark blue) represent the unique contributions of expression and functional optimality, respectively, and the three-way intersection (purple) represents the shared contribution of these two factors. For multicellular species, the expression in the tissue with the strongest ER for enzymes was used in the analysis (the brain basal ganglia for *H. sapiens* and the seedling root for *A. thaliana*).

may further increase the fraction of ER explained by functional optimality. Taken together, these results demonstrate not only an important role played by the optimization of protein efficiency in constraining protein evolution (Fig. 1), but also its major role in mediating the ER correlation.

Discussion

The presented results demonstrate that maintaining optimal protein function, for example high values of enzyme kinetic constants, imposes substantial constraints on protein sequences. This effect (the KR correlation, Fig. 1) significantly decreases the rate of amino acid substitutions and thus slows down protein evolution. Our analysis of empirical data shows that the variability in functional optimality across proteins explains a substantial fraction (30% to 40%) of the protein evolutionary rate variance in such diverse organisms as *H. sapiens*, *A. thaliana*, and *E. coli*. We note that in addition to the overall protein functional optimization other function- and structure-specific factors are likely to influence the rate of protein evolution (Wolf et al. 2008, 2010).

Our results suggest that a functional model of protein evolution (Fig. 8), which is based on the KR correlation and the FORCE mechanisms, may explain up to half of the ER correlation between the rate of protein evolution and protein expression. In addition to the KR correlation, we found that protein expression and functional efficiency also correlate with each other (the EK correlation, Fig. 6, a to c). The EK correlation likely emerges because both protein efficiency and expression level tend to increase together to meet the demand for the total protein activity in the cell. Because protein expression is usually associated with certain fitness costs (Dong et al. 1995; Dekel and Alon 2005; Plata et al. 2010; Scott et al. 2010; Kafri et al. 2016), increasing the total protein activity exclusively through upregulation of protein expression is disadvantageous. On the other hand, the functional optimization of protein sequence is limited by the entropic factor, i.e. there are many more sequences with sub-optimal than with optimal function. Balancing between the expression cost and the mutation-selection balance for the functional optimization, the cell satisfies the protein activity demand via both avenues simultaneously. As a result, proteins with high cellular activity demand tend to have both high expression and high functional efficiency, while proteins with low demand tend to have low expression and low functional efficiency. Because protein expression positively correlates with functional optimality (the EK correlation), and functional optimality decreases the rate of protein evolution (the KR correlation), highly expressed proteins usually evolve slowly, i.e. display the ER correlation (Fig. 8). Our results also show that protein functional optimality and evolutionary rate are primarily affected by expression in the same tissues and cell types, likely to be the ones most sensitive to expression costs (Fig. 6, d and e).

We note that the functional model of protein evolution (Fig. 8) is consistent with different origins of the protein expression costs and their various combinations (Cherry 2010). The key and unique component of the FORCE mechanism is not the existence of an expression cost, which is also a requirement of other evolutionary models (Drummond and Wilke 2008), but an essential role played by the optimization of protein functional efficiency that constrains protein sequence and slows protein evolution. As described above, the total protein activity in the cell can be increased by either the optimization of protein efficiency or by upregulation of protein expression (Parsch et al. 2000), and there is usually a saturating relationship between the total protein activity in the cell and species' fitness (Kacser and Burns 1981; Hartl et al. 1985). Consistent with the FORCE mechanism, previous experimental studies demonstrated that many coding

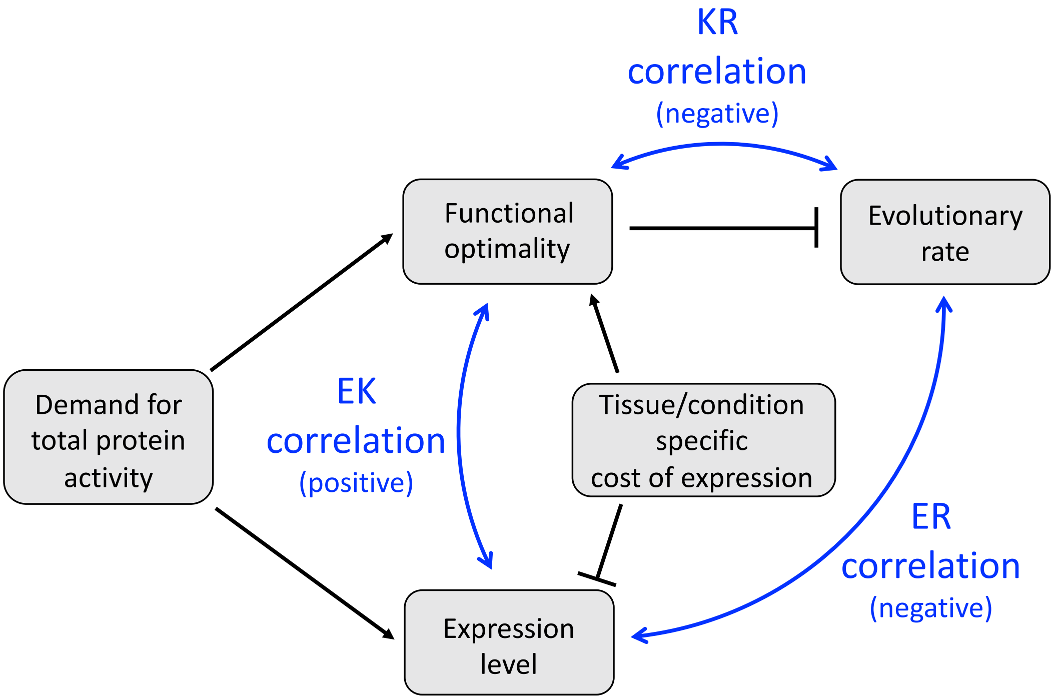


Fig. 8. Functional model of protein evolution. The figure illustrates the functional model of protein evolution and the Functional-Optimization-to-Reduce-the-Cost-of-Expression (FORCE) mechanism that are supported by the presented data. Gray rectangles represent molecular and cellular properties and processes. Black lines indicate stimulatory (arrows) and inhibitory (T-bars) effects. The bidirectional blue arrows indicate the key experimentally observed correlations: the negative ER correlation between expression and evolutionary rate, the negative KR correlation between protein functional optimality and evolutionary rate, and the positive EK correlation between expression and protein functional optimality. A higher demand for the total protein activity simultaneously leads to higher protein expression and increased functional optimality (the EK correlation). The requirement to maintain high functional optimality, in turn, constrains protein sequence and slows the rate of protein evolution (the ER correlation). Furthermore, particularly high costs of protein expression in certain tissues and conditions facilitate the optimization of protein functions and lead to stronger EK and ER correlations in the tissues most sensitive to protein expression costs.

mutations in enzymes have only small fitness effects when protein abundance is high, as a mutation-induced decrease in functional efficiency can be buffered when the total protein activity is close to saturation (Jiang et al. 2013; Wu et al. 2022; Cisneros et al. 2023). But the same coding mutations may significantly affect fitness when protein abundance is low, and thus protein optimality and protein expression level are both important for maintaining the total protein activity in the cell.

We find that in multicellular organisms, the strength of the ER correlation and the pressure to optimize protein function are usually associated with certain cell-specific processes, such as synaptic activities in animals (Fig. 3) and growth-related processes in plants (Fig. 5). Functional optimization may help to relieve the expression burden in plants' tissues with rapid cellular growth and active translation. Fast-growing plant cells experience a high demand for ATP and carbon required for biosynthesis. For example, it was estimated that in *A. thaliana* fast-growing leaves spend ∼40% of their ATP on protein production, while slow-growing leaves spend ∼3 times less (Li et al. 2017). Similarly, energetic and morphological properties of animal neurons, especially neurons with upregulated synaptic activities, are likely to result in particularly high costs of protein expression. Multiple evidence suggest that the brain and neurons are highly sensitive to energy limitations. The brain is known to oxidize glucose almost completely (Mergenthaler et al. 2013), and the glucose uptake per unit mass in the brain is two times higher than in other tissues (Lu et al. 2022). Furthermore, the majority (up to ∼80%) of neuronal ATP is used for synaptic repolarization (Alle et al. 2009; Harris et al. 2012; Magistretti and Allaman 2015). In addition to substantial energy consumption, large volumes of neuronal dendritic trees may provide a protein trafficking burden as protein expression primarily takes place in the soma (Maday et al. 2014). While one mechanism to reduce expression costs in neurons is to optimize protein function, another general mechanism is to decrease the rate of protein turnover. Indeed, the rate of protein turnover is substantially slower in the brain compared to other tissues (Fornasiero et al. 2018), and it is slower in neurons compared to glial cells (Dörrbaum et al. 2018). Slower evolution of proteins highly expressed in the brain may also lead to slower evolution of other cellular systems and properties. For example, it has been demonstrated that cellular transcriptome (Brawand et al. 2011; Chen et al. 2019), metabolome (Ma et al. 2015), and tissue-specific codon usage (Plotkin et al. 2004) also evolve significantly slower in the brain compared to other tissues.

Proteins, cells, and tissues of multicellular organisms do not function in isolation, but rather as an integrated system that insures proper physiological responses and species' survival. Therefore, it is of keen interest to understand the optimization of individual biological components, such as proteins, in the context of complex biological systems. We and others have previously investigated the influence of cellular protein–protein and metabolic networks on the evolution of individual proteins (Fraser et al. 2002; Vitkup et al. 2006). Our present work reveals the striking variability of protein functional optimization and explains how that variability affects protein evolution. We hope that our study will be an important step toward the development of an integrated functional theory of protein evolution which will jointly consider the effects associated with protein functional optimization and structural adaptation, expression patterns across tissues and conditions, and population demographic history.

Methods

Gene Expression Datasets Used in the Analyses

The following transcriptomes were used in this study: tissue-specific transcriptomes of *Homo sapiens* (Mele et al. 2015), *Mus musculus* (Söllner et al. 2017), *Drosophila melanogaster* (Leader et al. 2018), *Caenorhabditis elegans* (Spencer et al. 2011), *Arabidopsis thaliana* (Klepikova et al. 2016), *Zea mays* (Stelpflug et al. 2016) and *Glycine max* (Shen et al. 2014); cell-type-specific transcriptomes of the brain of *M. musculus* (Saunders et al. 2018; Zeisel et al. 2018; Sugino et al. 2019) and *D. melanogaster* (Davie et al. 2018); region-specific transcriptome of the *M. musculus* brain (Lein et al. 2007); tissue-specific transcriptome at different developmental stages of *M. musculus* (Cardoso-Moreira et al. 2019), and the transcriptome of *Escherichia coli* measured in log phase growth (McClure et al. 2013). Data sources, number of samples, sequencing technique, and specific details on data extraction for each dataset are provided in supplementary table S5, Supplementary Material online. In our analyses, we only used expression levels for the chromosomal protein-coding genes. For all single-cell datasets, we also excluded cell-type clusters with low expression resolution and used only clusters with at least 200,000 UMI counts. We applied the transcript per million (TPM) normalization for the tissue-specific transcriptomes of *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *E. coli* before subsequent analyses. We used the trimmed mean of M-values (TMM) normalization method (Robinson and Oshlack 2010) available in the edgeR package (Robinson et al. 2010) for all plant tissue-specific transcriptomes and for all animal single-cell transcriptomes to allow comparisons across samples in such analyses as the GSEA.

Calculation of Evolutionary Rates

We used the rate of non-synonymous substitutions, *dN*, as a measure of protein evolutionary rate. To calculate *dN* values for pairs of orthologous proteins, we utilized the PAML package (Yang 1997). We identified orthologous proteins as bidirectional best hits in pairwise local alignments between proteins from two species. The pairwise alignments were generated using Usearch (Edgar 2010), and included only protein pairs for which the corresponding alignments had *E*-value $<10^{-6}$, were at least 30 amino acids long, and covered at least 70% of the length of both proteins. The orthologous pairs of species used in our analysis were: *Mus musculus*—*Homo sapiens*, *Drosophila melanogaster*—*Drosophila yakuba*, *Caenorhabditis elegans*—*Caenorhabditis briggsae*, *Zea mays* (corn)—*Oryza sativa* (rice), *Glycine max* (soybean)—*Medicago truncatula*, and *Escherichia coli*—*Salmonella enterica*. The coding sequences for the proteins in each species were obtained from the Ensembl database (Cunningham et al. 2022).

For *H. sapiens* and *A. thaliana*, we used a multi-species approach to obtain accurate estimates of protein evolutionary rate. Notably, the standard procedure which uses only a pair of closely related species did not find reliable orthologs for a fraction of proteins, thus precluding us from inferring their evolutionary rates. To increase the number of orthologs in the analysis, we calculated the average evolutionary rate for *H. sapiens* and *A. thaliana* (primary species) proteins based on *dN* values obtained from multiple pairs of evolutionary related species, while also allowing protein orthologs in some species to be missing. Specifically, we used *Gorilla gorilla*, *Pongo abelii*, *Macaa mulatta*, *Saimiri boliviensis boliviensis*, and *Mus musculus* as the secondary species for *Homo sapiens*; and *Arabidopsis halleri*, *Brassica oleracea* (cabbage), *Glycine max* (soybean), *Solanum lycopersicum* (tomato), and *Helianthus annuus* (sunflower) as the secondary species for *Arabidopsis thaliana*.

First, we estimated the evolutionary length of the branches between the primary species, *H. sapiens* and *A. thaliana*, and their corresponding secondary species. To do this, we constructed a set of proteins from the primary species that have orthologs in all secondary species. Based on these sets, we calculated the average evolutionary distance between the primary and secondary species, *k*, as $\overline{dN_k} = \frac{\sum_i dN_k^i}{number\_of\_proteins}$, where dN_k^i is the non-synonymous evolutionary rate of the protein *i* calculated relative to the secondary species *k*, and the sum is over all proteins, *i*, in the set. Next, for each secondary species *k*, we calculated the relative evolutionary branch length $\alpha_k = \frac{\overline{dN_k}}{\sum_q \overline{dN_q}}$, where the sum is over all secondary species, *q*. Finally, for each protein *i* with orthologs in at least one secondary species, we estimated the mean protein-specific evolutionary rate as $\overline{dN^i} = \frac{\sum_k dN_k^i}{\sum_k \alpha_k}$, where the sum in the numerator and denominator is over the secondary species, *k*, that have orthologs of the protein *i*. The resulting protein-specific rates for orthologs in the primary species represent the fraction of non-synonymous substitutions along all evolutionary branches to the secondary species, normalized by the relative evolutionary length of the branches for which orthologs were detected.

The usage of the mean protein-specific rates, $\overline{dN}^i$, for *H. sapiens* and *A. thaliana* toward multiple secondary species increased by ∼15% to 20% the number of proteins with estimated evolutionary rates. Specifically, $\overline{dN}^i$ was inferred for 18,634 human and 23,076 *A. thaliana* proteins, compared to 16,846 and 19,042 proteins, respectively, when using only *Mus musculus* and *Brassica oleracea* as orthologous species (Zhang and Yang 2015).

To quantify the strength of purifying selection, we used the ratio of non-synonymous to synonymous substitution rates, *dN*/*dS*. To that end, for each pair of primary and secondary species, we calculated *dN*/*dS* values in the same way as *dN*, i.e. using orthologous proteins and the PAML package (Yang 1997). Since *dN*/*dS* values should not depend on the lengths of evolutionary branches and are similar for different secondary species, we used the analysis of the median value for each gene across different pairs of primary and secondary species.

Estimated evolutionary rates for each species are available in supplementary table S6, Supplementary Material online.

The polymorphism rate for human proteins was calculated as the number of non-synonymous protein-coding SNPs with frequencies greater than 1% in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) normalized by the protein length. The human polymorphism data was obtained from the dbSNP database (Sherry et al. 2001) (https://www.ncbi.nlm.nih.gov/snp/).

Expression—Evolutionary Rate Correlation (ER)

In the manuscript, we calculated the Expression—evolutionary Rate correlation (ER) as Spearman's correlation coefficient between mRNA expression levels and protein evolutionary rates. We used Spearman's correlation to avoid making any assumptions about the shape of the relationship between evolutionary rates and expression. Additionally, Spearman's correlation ensures that ER is invariant with respect to normalization and log-transformation of expression and evolutionary rate data. We note that all ER correlations calculated in this work are negative, indicating an anticorrelation between expression and evolutionary rate. However, for visualization purposes, we primarily use the correlation strength, quantified as the absolute value of the ER correlation, |ER|. Cell-type and tissue-specific ER correlations for each considered dataset are available in supplementary table S2, Supplementary Material online.

Linear Regression Analysis of Multi-tissue Contribution to ER

The relationship between expression and evolutionary rate is non-linear (Fig. 2a, supplementary fig. S3, a to c, Supplementary Material online). Therefore, to explore the joint influence of expression profiles across multiple tissues on evolutionary rates without making any assumptions about the shape of the ER relationship, we used rank-transformed expression values and rank-transformed evolutionary rates in the linear regression analysis. We then fitted a multivariable linear regression model based on expression in all tissues and compared its predictive power with the regression model based on expression in neural tissues only. The explanatory power of these regression models for different species is available in supplementary table S1, Supplementary Material online.

Expression Breadth Across Tissues

The breadth of expression for a gene was defined as the number of tissues in which it was expressed above a certain threshold (Park and Choi 2010). We explored several possible thresholds, namely 0, 0.1, 0.3, 1, 3, 10, 30, 100, 300, and 1,000 TPM. We then selected, for each animal species in the analysis, the threshold that provided the strongest Spearman's correlation between evolutionary rate and expression breadth; the selected threshold was 10 TPM for *H. sapiens*, *M. musculus*, and *D. melanogaster*, and 30 TPM for *C. elegans*. The Spearman's correlations between evolutionary rate and expression breadth are shown in supplementary table S1, Supplementary Material online.

Influence of Tissue-specific Genes on the ER Correlation

We defined the specificity of a particular gene to a tissue *k* as the z-score of its expression in tissue *k* relative to other tissues, $z = \frac{x-\mu}{\sigma+epsilon}$, where *x* is the expression level of the gene in tissue *k*, μ and σ are the mean and standard deviation of the expression level of the gene in other tissues, and *epsilon* is a parameter equal to the minimum non-zero expression level in the transcriptome. Expression values were log10(x + 1) transformed before the analysis. To investigate how well the expression of genes that are specific to a given tissue correlates with evolutionary rates, we selected various fractions of genes (ranging from 10% to 100%) with the highest specificity score to a given tissue and then used these genes to calculate the ER correlation for all tissues. This analysis was repeated for genes specific to each tissue (supplementary figs. S5 and S6, Supplementary Material online).

We used a similar approach to calculate the overall gene specificity to neural tissues in animals and to growing tissues in plants. In this case, the gene specificity score was defined as $z = \frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2+epsilon}}$, where μ_1 and σ_1^2 are the mean and variance of the gene expression in neural tissues of animals or in fast-growing plant tissues, μ_2 and σ_2^2 are the mean and variance of the gene expression in other tissues, n_1 and n_2 are the number of neural and non-neural tissues in animals or the number of fast- and slow-growing tissues in plants, and *epsilon* is a parameter with a value equal to the minimum non-zero gene expression level in the transcriptome. Expression values were log10(x + 1) transformed before the analysis. Since in plants the distinction between fast- and slow-growing tissues is not binary, we classified 20% of the tissues with the highest expression of growth markers as fast-growing and 20% of the tissues with the lowest expression of growth markers as slow-growing. To evaluate the influence of neural-specific genes in animals and fast-growth-specific genes in plants on the ER correlation, we removed 10% of the most neural- or fast-growth genes from the analysis and recalculated the ER correlations (supplementary table S2, Supplementary Material online).

Influence of the Number of Expressed Genes on the ER Correlations

The number of expressed genes varies across tissues. To explore how this variability might contribute to the differences in the strength of tissue-specific ER correlations we equalized the number of expressed genes across tissues. To that end, for each multi-tissue transcriptome, we identified the tissue with the minimum number of non-zero expressed genes, *n*. For each other tissue, we then sorted the genes based on their expression levels and selected *n* genes, starting with the most highly expressed, and set expression values for the other genes to zero. We then recalculated the ER correlations across tissues using these modified expression profiles and compared them to the original ER correlations (supplementary table S2, Supplementary Material online).

Expression of Essential Genes Across Tissues

We used the data describing the fitness effects of ∼7,000 knockouts for mouse genes obtained from the International Mouse Phenotyping Consortium (IMPC) database, available for download on 2022 April 18 [Koscielny et al. 2014 (https://www.mousephenotype.org/about-impc/)]. We defined genes as essential if their knockout viability phenotypes were classified in the database as "lethal." We calculated the fraction of essential genes expressed (at the level ≥1 TPM) in the mouse brain at different developmental stages; the results were not sensitive to the selected expression threshold (supplementary table S2, Supplementary Material online).

Gene Set Enrichment Analysis (GSEA)

Gene ontology (GO) annotations were obtained from the Molecular Signatures Database MSigDB:C5 v7.4 (release date 2021 March; Liberzon et al. 2011) for mouse genes, from the Arabidopsis Information Resource (TAIR) database (release date 2021 April 1; Berardini et al. 2015) for Arabidopsis genes, and from the AmiGo browser (release date 2021 February 1; Carbon et al. 2009) for fly, corn and soybean genes.

For GSEA analyses of single-cell transcriptomes, we maintained consistency in the resolution of expression profiles across cell-type clusters by retaining the same number of non-zero expressed genes in each cell-type cluster. To achieve this, we sorted genes by their expression for each cell type and retained the top 10,000 most highly expressed genes for mouse and 7,000 genes for *D. melanogaster*. Expression of other genes was set to zero. We note that GSEA analyses performed on uncorrected expression data yielded very similar sets of upregulated GO terms (supplementary table S4, Supplementary Material online).

We used the gene set enrichment analysis to identify the functional roles of genes with upregulated expression in cell types exhibiting stronger ER correlations. To this end, we first calculated, for each gene, the Pearson's correlation coefficient between its expression across different cell-type clusters and the strength of the cell-type-specific ER, $r(Expression,\ |ER|)$. We then ranked the genes based on the strength of this correlation and performed a pre-ranked GSEA analysis (Subramanian et al. 2005), using the "prerank" module of GSEAPY (Fang et al. 2023), to identify GO terms enriched among genes associated with stronger ER. Following the default settings, we used in the GSEA analysis GO terms containing ≥15, but ≤2,000 genes.

To verify the robustness of our results, we also utilized an alternative procedure to identify the association between gene expression and ER strength. Specifically, we sorted cell types by the strength of ER and calculated, for each gene, the difference in its average expression levels between the top 10% and bottom 90% (or top 50% and bottom 50%) of the sorted cell types. We then performed a pre-ranked GSEA analysis based on the calculated differential expression values.

We further investigated the GSEA results in mouse to understand whether the association between the ER strength and expression of genes from synaptic and other enriched GO terms was due to the direct influence of these genes on the ER correlation. To address this question, we removed all genes annotated with (i) the Synapse GO term (GO:0045202) or (ii) any of the significantly enriched GO terms (at $FDR < 0.05$). For each cell-type, we then recalculated the Spearman's correlation between evolutionary rates and expression using the remaining genes, $ER_{noSynapse}$ and $ER_{noEnriched}$, respectively. Next, we repeated the pre-ranked GSEA analysis, but this time the gene ranking (for all genes, including synaptic and associated with other enriched GO terms) was based on Pearson's correlation between gene expression and the strength of $ER_{noSynapse}$ or $ER_{noEnriched}$, $r(Expression,\ |ER_{noSynapse}|)$ or $r(Expression,\ |ER_{noEnriched}|)$, respectively.

The results of the GSEA analyses for all animals' and plants' datasets and corresponding controls are available in supplementary table S4, Supplementary Material online.

Hierarchical Clustering of Expression Samples

We performed hierarchical clustering of expression samples for each plant species using the "hclust" function from the R package "stats" (R Core Team 2022). The distance between samples was calculated as one minus the squared Pearson's correlation coefficient between corresponding expression profiles. We used the "ward.D2" agglomeration method for corn and the "complete" method for Arabidopsis and soybean.

Markers for Growth Stages in Plants

Marker genes for the division, elongation, and differentiation growth stages in Arabidopsis were taken from Huang and Schiefelbein (2015). For corn and soybean, we used orthologs of the Arabidopsis marker genes as markers for the corresponding growth stages. Ortholog annotations were obtained from the Ensembl database (Cunningham et al. 2022) via BioMart (Kinsella et al. 2011), using the "one2any" homology and confidence level "1". The average expression of genes from each growth stage was calculated after log10(x + 1) transformation of expression values. The final lists of marker genes for each plant species are available in supplementary table S7, Supplementary Material online.

Collection of Data on Enzymes' Catalytic Rates

We extracted all available data on k_{cat} and k_{cat}/K_M from the Brenda (Chang et al. 2021) (version of 2019 September 3) and Sabio-RK (Wittig et al. 2018) (version of 2019 January 21) databases. Specifically, we downloaded from Brenda (https://www.brenda-enzymes.org/download.php) an easy-to-parse text file containing catalytic rates along with additional information about the entries, such as reaction EC numbers, protein Uniprot IDs, protein names and types, source organisms, and measurement temperatures. For Sabio-RK, we used the provided URL request interface to automatically download all kinetic constants and entries' information. In addition, we obtained data on k_{cat} for *E. coli* enzymes from the previously published and manually curated dataset (Davidi et al. 2016). We then applied several filters to exclude mutant enzymes, enzymes with macromolecular substrates or those involved in transmembrane transport, and multifunctional enzymes that catalyze multiple reactions with EC numbers differing by more than the last digit. In cases where multiple values of k_{cat} or k_{cat}/K_M were available for a given enzyme, for example, due to measurements with different substrates or measurements available in different publications, we used the highest values of experimentally obtained kinetic constants.

For the final sets of enzymes from *H. sapiens*, *A. thaliana*, and *E. coli*, we manually curated the references where k_{cat}, k_{cat}/K_M, and corresponding k_{cat}^{max}, and $(k_{cat}/K_m)^{max}$ (described in the next section) were published. We observed some discrepancies between the values given in the original publications and the corresponding entries in the Brenda or Sabio-RK databases. The most common errors were due to incorrect transfer of units, for example, the usage of mM instead of μM or s^{-1} instead of min^{-1}. All such cases were corrected in our dataset.

The complete dataset of catalytic rates for the enzymes used in this work is available in supplementary table S8, Supplementary Material online. Data on catalytic rates for enzymes from *H. sapiens*, *A. thaliana*, and *E. coli*, combined with their evolutionary rates and expression values are available in supplementary table S9, Supplementary Material online.

Normalization of Catalytic Rates to Estimate Functional Optimization

Following the previous approach (Davidi et al. 2018), to estimate the level of functional optimization of the considered enzymes we calculated the normalized catalytic rates, k_{cat}^{norm} or $(k_{cat}/K_M)^{norm}$. To that end, we divided the kinetic constants, k_{cat} and k_{cat}/K_M, by the highest constants known for the corresponding reaction classes, k_{cat}^{max} and $(k_{cat}/K_M)^{max}$. This normalization procedure allowed us to account for the substantial heterogeneity of kinetic constant values across different reaction classes due to the diverse chemistries of the catalyzed reactions. To perform the normalization, we identified, for each reaction class (i.e. enzymes sharing all four digits of the EC classification), the highest catalytic rates measured across all database entries (including mutants and multifunctional enzymes), k_{cat}^{max} or $(k_{cat}/K_M)^{max}$. To ensure accurate estimates of the maximum catalytic rate for a given reaction class, we included in the analysis only enzymes for which experimental measurements of catalytic rates were available for at least 15 unique enzymes in the same reaction class (EC number); the results were not very sensitive to the value of this parameter.

Temperature Correction of the Measured Catalytic Rates

The majority of the catalytic rate constants in our dataset were measured at temperatures in the range of 20 to 40 °C, with the median at ∼30 °C. However, approximately 6% of the available kinetic measurements were performed at higher temperatures, up to 100 °C. Since the rates of catalyzed biochemical reactions are temperature-dependent according to Arrhenius' law, we applied a temperature correction to estimate the catalytic rates of the corresponding enzymes at 30 °C. We note that the exact dependence of the reaction rate on temperature depends on the activation energy, which is specific to each catalyzed chemical reaction. However, a previous systematic analysis of temperature-dependent acceleration of biochemical reactions, catalyzed by more than 150 enzymes, demonstrated (Elias et al. 2014) that, on average, the values of k_{cat} and k_{cat}/K_M increased by a factor of 1.8 for each 10 °C increase in temperature. To account for this trend, we scaled down all k_{cat} and k_{cat}/K_M values experimentally measured at temperatures $T > 40$ °C using the following equation: $k(30^{\circ}C) = k(T) \cdot 1.8^{(T-30)/10}$, where k is the catalytic constant and T is the measurement temperature; we did not apply any adjustments to the catalytic rate constants measured at $T < 40$ °C. In cases where the measurement temperature was not explicitly stated in the publication, we assumed that the measurements were performed at ambient temperature and did not apply any temperature correction.

We note that temperature correction only affected the functional efficiency of 21 enzymes (9%) in the final sets for *H. sapiens*, *A. thaliana*, and *E. coli*. Moreover, functional efficiencies calculated based on the uncorrected data showed correlations with protein evolutionary rates and expression that were very similar to those observed in the temperature-corrected data (supplementary table S10, Supplementary Material online).

Multi-member EC Classes

Species' genomes often encode several distinct enzymes that catalyze the same EC reaction. Among the sets of enzymes with estimated functional efficiency, most reaction classes were represented by only one or two different proteins. However, for *H. sapiens* and *A. thaliana*, several EC classes were represented by multiple enzymes. These enzymes usually had different expression levels, evolutionary rates, and catalytic efficiencies (supplementary table S9, Supplementary Material online). However, to assess the potential influence of multi-member EC classes on the KR correlation, we randomly subsampled the enzymes from the multi-member EC classes, retaining only two different enzymes from each EC class with more than two protein members. For each of the 10,000 such random trials, we recalculated the KR correlation and the fraction of the ER correlation mediated by the KR correlation. Although this procedure reduced the number of enzymes by about 1/3, the KR correlation was significant in more than 99% of the reduced samples, and the median KR correlation coefficients and the fractions of ER explained by functional efficiency were similar to the results obtained from the complete enzyme sets (supplementary table S10, Supplementary Material online).

Unique and Shared Contributions of Protein Optimality and Expression to Explaining Protein Evolutionary Rates

We used semi-partial correlations (Abdi 2007) to investigate the unique and shared contributions of two independent variables, i.e. protein functional optimality and expression, to explain the variance of the dependent variable, protein evolutionary rates. To do this, the unique effect of one independent variable x on the dependent variable z was calculated as the squared coefficient of the semi-partial correlation between them while controlling for the other independent variable, y: $r^2_{z(x.y)}$. Similarly, the unique effect of the independent variable y on the dependent variable z was calculated as $r^2_{z(y.x)}$. The shared effect was calculated as the difference between the squared coefficient of the ordinary bivariate correlation, $r^2_{z.x}$, and the squared coefficient of the semi-partial correlation $r^2_{z(x.y)}$. We used the function "partial_cor" with the method "spearman" from the Python package "pingouin" to calculate semi-partial correlations (Vallat 2018).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Dr. Eugene Koonin, Dr. Yuri Wolf, Dr. Aris Floratos, Dr. Olivier Tenaillon, Dr. Balázs Papp, and the members of the Vitkup lab for many helpful and stimulating scientific discussions. This work was supported in part by the National Institute of General Medical Sciences Grant No. R35GM131884 and the National Institute of Mental Health Grant No. 5R01MH124923 to D.V.

Conflict of interest: All authors declare no competing interests.

Data availability

The data supporting the findings of this article are provided within the main text and the supplementary materials.

References

Abdi H. Part (semi partial) and partial regression coefficients. In: Salkind NJ, editor. Encyclopedia of measurement and statistics. Thousand Oaks, CA, USA: Sage Publications; 2007. Vol 3. p. 736–739.

Abeles RH, Frey PA, Jencks WP. Biochemistry. Boston, MA, USA: Jones and Bartlett; 1992.

Alle H, Roth A, Geiger JR. Energy-efficient action potentials in hippocampal mossy fibers. Science. 2009:**325**(5946):1405–1408. https://doi.org/10.1126/science.1174331.

Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011:**50**(21):4402–4410. https://doi.org/10.1021/bi2002289.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. Genesis. 2015:**53**(8):474–485. https://doi.org/10.1002/dvg.22877.

Biesiadecka MK, Sliwa P, Tomala K, Korona R. An overexpression experiment does not support the hypothesis that avoidance of toxicity determines the rate of protein evolution. Genome Biol Evol. 2020:**12**(5):589–596. https://doi.org/10.1093/gbe/evaa067.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011:**478**(7369):343–348. https://doi.org/10.1038/nature10532.

Buttgereit F, Brand MD. A hierarchy of ATP-consuming processes in mammalian cells. Biochem J. 1995:**312**(1):163–167. https://doi.org/10.1042/bj3120163.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S; AmiGO Hub; Web Presence Working Group. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009:**25**(2):288–289. https://doi.org/10.1093/bioinformatics/btn615.

Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, Liechti A, Ascenção K, Rummel C, Ovchinnikova S, et al. Gene expression across mammalian organ development. Nature. 2019:**571**(7766):505–509. https://doi.org/10.1038/s41586-019-1338-5.

Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021:**49**(D1):D498–D508. https://doi.org/10.1093/nar/gkaa1025.

Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, Haerty W, Palma FD, Regev A. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res. 2019:**29**(1):53–63. https://doi.org/10.1101/gr.237636.118.

Chen L, Vitkup D. Distribution of orphan metabolic activities. Trends Biotechnol. 2007:**25**(8):343–348. https://doi.org/10.1016/j.tibtech.2007.06.001.

Cherry JL. Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol. 2010:**2**:757–769. https://doi.org/10.1093/gbe/evq059.

Cisneros AF, Gagnon-Arsenault I, Dubé AK, Després PC, Kumar P, Lafontaine K, Pelletier JN, Landry CR. Epistasis between promoter activity and coding mutations shapes gene evolvability. Sci Adv. 2023:**9**(5):eadd9109. https://doi.org/10.1126/sciadv.add9109.

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R. Ensembl 2022. Nucleic Acids Res. 2022:**50**(D1):D988–D995. https://doi.org/10.1093/nar/gkab1049.

Davidi D, Longo LM, Jablonska J, Milo R, Tawfik DS. A bird's-eye view of enzyme evolution: chemical, physicochemical, and physiological considerations. Chem Rev. 2018:**118**(18):8786–8797. https://doi.org/10.1021/acs.chemrev.8b00039.

Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, Barenholz U, Goldenfeld M, Shlomi T, Milo R. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proc Natl Acad Sci U S A. 2016:**113**(12):3401–3406. https://doi.org/10.1073/pnas.1514240113.

Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft Ł, Aibar S, Makhzami S, Christiaens V, Bravo González-Blas C, et al. A single-cell transcriptome atlas of the aging Drosophila brain. Cell. 2018:**174**(4):982–998.e20. https://doi.org/10.1016/j.cell.2018.05.057.

Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005:**436**(7050):588–592. https://doi.org/10.1038/nature03842.

Dickerson RE. The structures of cytochrome c and the rates of molecular evolution. J Mol Evol. 1971:**1**(1):26–45. https://doi.org/10.1007/BF01659392.

Dong H, Nilsson L, Kurland CG. Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. J Bacteriol. 1995:**177**(6):1497–1504. https://doi.org/10.1128/jb.177.6.1497-1504.1995.

Dörrbaum AR, Kochen L, Langer JD, Schuman EM. Local and global influences on protein turnover in neurons and glia. eLife. 2018:**7**:e34202. https://doi.org/10.7554/eLife.34202.

Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008:**134**(2):341–352. https://doi.org/10.1016/j.cell.2008.05.042.

Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 2000:**17**(1):68–74. https://doi.org/10.1093/oxfordjournals.molbev.a026239.

Echave J. Beyond stability constraints: a biophysical model of enzyme evolution with selection on stability and activity. Mol Biol Evol. 2019:**36**(3):613–620. https://doi.org/10.1093/molbev/msy244.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010:**26**(19):2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

Elias M, Wieczorek G, Rosenne S, Tawfik DS. The universality of enzymatic rate–temperature dependency. Trends Biochem Sci. 2014:**39**(1):1–7. https://doi.org/10.1016/j.tibs.2013.11.001.

Erö C, Gewaltig MO, Keller D, Markram H. A cell atlas for the mouse brain. Front Neuroinform. 2018:**12**:84. https://doi.org/10.3389/fninf.2018.00084.

Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023:**39**(1):btac757. https://doi.org/10.1093/bioinformatics/btac757.

Ferreiro D, Khalil R, Sousa SF, Arenas M. Substitution models of protein evolution with selection on enzymatic activity. Mol Biol Evol. 2024:**41**(2):msae026. https://doi.org/10.1093/molbev/msae026.

Feugeas J-P, Tourret J, Launay A, Bouvet O, Hoede C, Denamur E, Tenaillon O. Links between transcription, environmental adaptation and gene variability in *Escherichia coli*: correlations between gene expression and gene variability reflect growth efficiencies. Mol Biol Evol. 2016:**33**(10):2515–2529. https://doi.org/10.1093/molbev/msw105.

Fornasiero EF, Mandad S, Wildhagen H, Alevra M, Rammner B, Keihani S, Opazo F, Urban I, Ischebeck T, Sakib MS, et al. Precisely measured protein lifetimes in the mouse brain reveal differences across tissues and subcellular fractions. Nat Commun. 2018:**9**(1):4230. https://doi.org/10.1038/s41467-018-06519-0.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002:**296**(5568):750–752. https://doi.org/10.1126/science.1068696.

Futuyma DJ, Kirkpatrick M. Evolution IV edition. Sunderland (MA): Sinauer Associates, Inc; 2017.

Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins. 2011:**79**(5):1396–1407. https://doi.org/10.1002/prot.22964.

Gout JF, Kahn D, Duret L; Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 2010:**6**(5):e1000944. https://doi.org/10.1371/journal.pgen.1000944.

Harris JJ, Jolivet R, Attwell D. Synaptic energy use and supply. Neuron. 2012:**75**(5):762–777. https://doi.org/10.1016/j.neuron.2012.08.019.

Hartl DL, Dykhuizen DE, Dean AM. Limits of adaptation: the evolution of selective neutrality. Genetics. 1985:**111**(3):655–674. https://doi.org/10.1093/genetics/111.3.655.

Hu G, Li J, Wang G-Z. Significant evolutionary constraints on neuron cells revealed by single-cell transcriptomics. Genome Biol Evol. 2020:**12**(4):300–308. https://doi.org/10.1093/gbe/evaa054.

Huang L, Schiefelbein J. Conserved gene expression programs in developing roots from diverse plants. Plant Cell. 2015:**27**(8):2119–2132. https://doi.org/10.1105/tpc.15.00328.

Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol. 2016:**14**(5):e1002452. https://doi.org/10.1371/journal.pbio.1002452.

Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA. Latent effects of Hsp90 mutants revealed at reduced expression levels. PLoS Genet. 2013:**9**(6):e1003600. https://doi.org/10.1371/journal.pgen.1003600.

Kacser H, Burns JA. The molecular basis of dominance. Genetics. 1981:**97**(3-4):639–666. https://doi.org/10.1093/genetics/97.3-4.639.

Kafri M, Metzl-Raz E, Jona G, Barkai N. The cost of protein production. Cell Rep. 2016:**14**(1):22–31. https://doi.org/10.1016/j.celrep.2015.12.015.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011:**2011**:bar030. https://doi.org/10.1093/database/bar030.

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant J. 2016:**88**(6):1058–1070. https://doi.org/10.1111/tpj.13312.

Konate MM, Plata G, Park J, Usmanova DR, Wang H, Vitkup D. Molecular function limits divergent protein evolution on planetary timescales. eLife. 2019:**8**:e39705. https://doi.org/10.7554/eLife.39705.

Koonin EV, Wolf YI. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 2006:**17**(5):481–487. https://doi.org/10.1016/j.copbio.2006.08.003.

Koonin EV, Wolf YI. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet. 2010:**11**(7): 487–498. https://doi.org/10.1038/nrg2810.

Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen C-K, Easty R, Di Fenza A, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res. 2014:**42**(D1):D802–D809. https://doi.org/10.1093/nar/gkt977.

Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic Acids Res. 2018:**46**(D1):D809–D815. https://doi.org/10.1093/nar/gkx976.

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007:**445**(7124):168–176. https://doi.org/10.1038/nature05453.

Li L, Nelson CJ, Trösch J, Castleden I, Huang S, Millar AH. Protein degradation rate in *Arabidopsis thaliana* leaf growth and development. Plant Cell. 2017:**29**(2):207–228. https://doi.org/10.1105/tpc.16.00768.

Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985:**2**(2):150–174. https://doi.org/10.1093/oxfordjournals.molbev.a040343.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, TamayoP, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011:**27**(12):1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

Lu W, Cheng Z, Xie X, Li K, Duan Y, Li M, Ma C, Liu S, Qiu J. An atlas of glucose uptake across the entire human body as measured by the total-body PET/CT scanner: a pilot study. Life Metab. 2022:**1**(2):190–199. https://doi.org/10.1093/lifemeta/loac030.

Ma S, Lee S-G, Kim EB, Park TJ, Seluanov A, Gorbunova V, Buffenstein R, Seravalli J, Gladyshev VN. Organization of the mammalian ionome according to organ origin, lineage specialization, and longevity. Cell Rep. 2015:**13**(7):1319–1326. https://doi.org/10.1016/j.celrep.2015.10.014.

Maday S, Twelvetrees AE, Moughamian AJ, Holzbaur EL. Axonal transport: cargo-specific mechanisms of motility and regulation. Neuron. 2014:**84**(2):292–309. https://doi.org/10.1016/j.neuron.2014.10.019.

Magistretti PJ, Allaman I. A cellular perspective on brain energy metabolism and functional imaging. Neuron. 2015:**86**(4):883–901. https://doi.org/10.1016/j.neuron.2015.03.035.

McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. 2013:**41**(14):e140.

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015:**348**(6235):660–665. https://doi.org/10.1126/science.aaa0355.

Mergenthaler P, Lindauer U, Dienel GA, Meisel A. Sugar for the brain: the role of glucose in physiological and pathological brain function. Trends Neurosci. 2013:**36**(10):587–597. https://doi.org/10.1016/j.tins.2013.07.001.

Murakami TC, Mano T, Saikawa S, Horiguchi SA, Shigeta D, Baba K, Sekiya H, Shimizu Y, Tanaka KF, Kiyonari H, et al. A three-dimensional single-cell-resolution whole-brain atlas using CUBIC-X expansion microscopy and tissue clearing. Nat Neurosci. 2018:**21**(4):625–637. https://doi.org/10.1038/s41593-018-0109-1.

Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics. 2001:**158**(2):927–931. https://doi.org/10.1093/genetics/158.2.927.

Pal C, Papp B, Lercher MJ. An integrated view of protein evolution. Nat Rev Genet. 2006:**7**(5):337–348. https://doi.org/10.1038/nrg1838.

Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol Biol. 2010:**10**:241. https://doi.org/10.1186/1471-2148-10-241.

Parsch J, Russell JA, Beerman I, Hartl DL, Stephan W. Deletion of a conserved regulatory element in the Drosophila Adh gene leads to increased alcohol dehydrogenase activity but also delays development. Genetics. 2000:**156**(1):219–227. https://doi.org/10.1093/genetics/156.1.219.

Plata G, Gottesman ME, Vitkup D. The rate of the molecular clock and the cost of gratuitous protein synthesis. Genome Biol. 2010:**11**(9):R98. https://doi.org/10.1186/gb-2010-11-9-r98.

Plata G, Vitkup D. Protein stability and avoidance of toxic misfolding do not explain the sequence constraints of highly expressed proteins. Mol Biol Evol. 2018:**35**(3):700–703. https://doi.org/10.1093/molbev/msx323.

Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A. 2004:**101**(34):12588–12591. https://doi.org/10.1073/pnas.0404957101.

R Core Team. R: A language and environment for statistical computing. 2022. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Robinson MD, McCarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010:**26**(1):139–140. https://doi.org/10.1093/bioinformatics/btp616.

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010:**11**(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rocha EP. The quest for the universals of protein evolution. Trends Genet. 2006:**22**(8):412–416. https://doi.org/10.1016/j.tig.2006.06.004.

Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 2004:**21**(1):108–116. https://doi.org/10.1093/molbev/msh004.

Rolfe D, Brown GC. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. Physiol Rev. 1997:**77**(3): 731–758. https://doi.org/10.1152/physrev.1997.77.3.731.

Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, et al. Molecular diversity and specializations among the cells of the adult mouse brain. Cell. 2018:**174**(4):1015–1030.e1016. https://doi.org/10.1016/j.cell.2018.07.028.

Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010:**330**(6007):1099–1102. https://doi.org/10.1126/science.1192588.

Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong L-A, Peng D-L, et al. Global dissection of alternative splicing in paleopolyploid soybean. Plant Cell. 2014:**26**(3):996–1008. https://doi.org/10.1105/tpc.114.122739.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001:**29**(1):308–311. https://doi.org/10.1093/nar/29.1.308.

Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, Simon E. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Sci Data. 2017:**4**(1):170185. https://doi.org/10.1038/sdata.2017.185.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. A spatial and temporal map of *C. elegans* gene expression. Genome Res. 2011:**21**(2):325–341. https://doi.org/10.1101/gr.114595.110.

Stelpflug SC, Sekhon RS, Vaillancourt B, Hirsch CN, Buell CR, de Leon N, Kaeppler SM. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development.

Plant Genome. 2016:**9**(1):1–16. https://doi.org/10.3835/plantgenome2015.04.0025.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005:**102**(43):15545–15550. https://doi.org/10.1073/pnas.0506580102.

Sugino K, Clark E, Schulmann A, Shima Y, Wang L, Hunt DL, Hooks BM, Trankner D, Chandrashekar J, Picard S, et al. Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. eLife. 2019:**8**:e38619. https://doi.org/10.7554/eLife.38619.

Taiz L, Zeiger E, Møller IM, Murphy A. Plant physiology and development. 6th ed. Sunderland, CT: Sinauer Associates; 2015.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015:**526**(7571):68–74. https://doi.org/10.1038/nature15393.

Tuller T, Kupiec M, Ruppin E. Evolutionary rate and gene expression across different brain regions. Genome Biol. 2008:**9**(9):R142. https://doi.org/10.1186/gb-2008-9-9-r142.

Usmanova DR, Plata G, Vitkup D. The relationship between the misfolding avoidance hypothesis and protein evolutionary rates in the light of empirical evidence. Genome Biol Evol. 2021:**13**(2):evab006. https://doi.org/10.1093/gbe/evab006.

Vallat R. Pingouin: statistics in Python. J Open Source Softw. 2018:**3**(31):1026. https://doi.org/10.21105/joss.01026.

Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 2006:**7**(5):R39. https://doi.org/10.1186/gb-2006-7-5-r39.

Wang H-Y, Chien H-C, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou C-K, Tsai S-F, Wu C-I, Shen C-KJ. Rate of evolution in brain-expressed genes in humans and other primates. PLoS Biol. 2007:**5**(2):e13. https://doi.org/10.1371/journal.pbio.0050013.

Webb EC. Enzyme nomenclature 1992. Recommendations of the nomenclature committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Academic Press; 1992.

Wittig U, Rey M, Weidemann A, Kania R, Müller W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res. 2018:**46**(D1):D656–D660. https://doi.org/10.1093/nar/gkx1065.

Wolf MY, Wolf YI, Koonin EV. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. Biol Direct. 2008:**3**:40. https://doi.org/10.1186/1745-6150-3-40.

Wolf YI, Gopich IV, Lipman DJ, Koonin EV. Relative contributions of intrinsic structural–functional constraints and translation rate to the evolution of protein-coding genes. Genome Biol Evol. 2010:**2**:190–199. https://doi.org/10.1093/gbe/evq010.

Wu Z, Cai X, Zhang X, Liu Y, Tian G-B, Yang J-R, Chen X. Expression level is a major modifier of the fitness landscape of a protein coding gene. Nat Ecol Evol. 2022:**6**(1):103–115. https://doi.org/10.1038/s41559-021-01578-x.

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997:**13**(5):555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, et al. Molecular architecture of the mouse nervous system. Cell. 2018:**174**(4):999–1014.e22. https://doi.org/10.1016/j.cell.2018.06.021.

Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015:**16**(7):409–420. https://doi.org/10.1038/nrg3950.

Zhang L, Li W-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 2004:**21**(2):236–239. https://doi.org/10.1093/molbev/msh010.

Zhou T, Drummond DA, Wilke CO. Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol. 2008:**66**(4):395–404. https://doi.org/10.1007/s00239-008-9094-4.

Zhu F, Cizeron M, Qiu Z, Benavides-Piccione R, Kopanitsa MV, Skene NG, Koniaris B, DeFelipe J, Fransén E, Komiyama NH, et al. Architecture of the mouse brain synaptome. Neuron. 2018:**99**(4):781–799.e10. https://doi.org/10.1016/j.neuron.2018.07.007.

Zuckerkandl E, Pauling L. Molecular disease, evolution, and genic heterogeneity. In: Horizons in biochemistry. New York: Academic Press; 1962. p. 189–225.

Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins. New York: Academic Press; 1965. p. 97–166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.