═══REVIEW COMMONS MANUSCRIPT═══

IMPORTANT:

- Manuscripts submitted to Review Commons are peer reviewed in a journal-agnostic way.
- Upon transfer of the peer reviewed preprint to a journal, the referee reports will be available in full to the handling editor.
- The identity of the referees will NOT be communicated to the authors unless the reviewers choose to sign their report.
- The identity of the referee will be confidentially disclosed to any affiliate journals to which the manuscript is transferred.

GUIDELINES:

- For reviewers: https://www.reviewcommons.org/reviewers
- For authors: https://www.reviewcommons.org/authors

CONTACT:

The Review Commons office can be contacted directly at: office@reviewcommons.org

Cost-saving, trading, and internalization of externality: three economic strategies underlying amino acid archetypes of human metabolic enzymes

Kaixiang Yang[1], Kai Sun[1], Hu Zhu[2,3,4], Yu Liu[2,4], Ziwei Dai[1*]
1 Department of Systems Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China.
2 Department of Systems Science, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China.
3 School of Systems Science, Beijing Normal University, Beijing 100875, China.
4 International Academic Center of Complex Systems, Beijing Normal University, Zhuhai 519087, China.

*Correspondence: daizw@sustech.edu.cn

Abstract

Amino acid composition of proteome reflects principles of metabolic resource allocation. While amino acids with higher biosynthetic costs are used less frequently in unicellular organisms, whether this rule extends to complex multicellular life remains elusive. Here we show that amino acid composition of human metabolic enzymes does not follow a simple cost-minimizing mode but instead segregate into four distinct amino-acid composition modules each associated with specific metabolic functions, a phenomenon we term emergence of amino acid archetypes (AArchetypes). By comparing ten phylogenetically diverse species, we find that AArchetypes emerge only in multicellular organisms, indicating a fundamental shift in proteome-level economics. We further demonstrate that these AArchetypes correspond to three distinct economic strategies: cost-saving in young, fast-evolving enzymes; trading in conserved central carbon metabolic enzymes whose amino-acid compositions match that of dietary intake; and internalization of negative externalities in liquid-liquid phase separation (LLPS)-prone, intrinsically disordered proteins through elevated biosynthetic costs of LLPS-

associated amino acids. Together, these results reveal that multicellular proteomes are governed not by a single cost-minimization principle, but by a multi-layered economic system that integrates production, exchange, and regulation.

Introduction

Amino acids, fundamental building blocks of proteins, are not only symbols for encoding protein function but also physical entities that must be paid for with carbon, energy, and enzymatic capacity. *De novo* synthesis of amino acids requires three resources: carbon and nitrogen sources that provide the molecular backbone, energy that facilitates overcoming free energy barriers, and metabolic enzymes that catalyze reactions. The need to save costs of amino acid synthesis creates a simple economic law governing the organization of proteome in unicellular organisms: expensive amino acids are rare, and cheap ones dominate the proteome. The trade-off between cost-saving and maximizing protein sequence diversity produces a nearly universal negative linear relationship between costs of amino acids and logarithms of their frequencies in proteomes across bacteria and unicellular eukaryotes such as *E.coli*, *B. subtilis*, and *S.cerevisiae*[1-4], as well as cancer cells[5]. This cost-frequency law has proven remarkably robust to environmental factors such as temperature[6], redox potential[7], and ocean depth[8] that are known to affect amino acid composition of proteome.

Nevertheless, this law becomes conceptually fragile in complex multicellular organisms such as humans. Unlike bacteria and unicellular eukaryote, humans cannot synthesize all standard amino acids and must rely on dietary intake. In human body, amino acids are obtained through diet, buffered by circulation, and redistributed across tissues with diverse metabolic and functional demands. Under such circumstances, the cost of *de novo* synthesis no longer reflects the true economic price of an amino acid. Dietary amino

acid compositions matching cellular proteome have beneficial outcomes in multicellular animals such as fruit fly[9, 10], implying an economic strategy of trading instead of production. Furthermore, the difference in amino acid synthesis pathways between multicellular and unicellular organisms[11] could also profoundly change the costs of *de novo* amino acid synthesis.

Furthermore, beyond the cost-frequency coupling of amino acids at the whole proteome level, a proteome is fundamentally not a single economic agent. Instead, it is a collection of proteins each serving distinct functions, operating in different cellular compartments, and facing different evolutionary and physiological constraints. Consequently, the cost-frequency law at proteome level does not constrain amino acid composition of single proteins. It operates at the protein abundance, where proteins with costly amino acid compositions have lower expression levels[12]. However, this observation does not address the important question of why such costly amino acid composition profiles still exist in a variety of proteins.

In this study, we focus on metabolic enzymes that represent a substantial fraction of cellular proteome to investigate principles governing the diversification of amino acid compositions of proteins. Here we show that, in multicellular organisms such as humans, the cost-frequency law does not simply weaken. Instead, it diverges into three distinct economic strategies. Four representative patterns of amino acid composition emerge in human metabolic enzymes, each enriched for specific biological functions and reflecting a specific strategy for managing metabolic cost. We refer to this phenomenon as emergence of amino acid archetypes (AArchetypes) and showed that AArchetypes only emerge in multicellular organisms.

By linking human AArchetypes to protein age, function, and phase-separation ability, we show that these AArchetypes correspond to three economic

strategies: first, cost-saving, adopted by fast-evolving, young proteins to minimize synthetic costs of amino acids; second, trading, whereby conserved central carbon metabolic enzymes align their amino acid composition with dietary amino acid profiles; third, internalization of externality, in which liquid-liquid phase separation (LLPS)-related amino acids are costlier, reflecting the possibility of “taxing” LLPS-related amino acids to prevent excessive condensate formation. These findings together reveal that amino acid composition of multicellular proteomes is shaped not only by a single cost-minimization principle, but by a layered economic system involving fundamentally different economic strategies.

Materials and methods

Amino acid composition of metabolic enzymes and pathways

Information of metabolic pathways (i.e. KEGG pathways under the category “Metabolism”) was retrieved from the KEGG Pathway database. For ten species including *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Escherichia coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis*, and *Saccharomyces cerevisiae*, species-specific lists of genes and proteins encoding metabolic enzymes in each KEGG metabolic pathway and amino acid sequences of these enzymes were retrieved using the KEGG API. Relative frequency of each of the 20 standard amino acids across the full-length sequence (i.e., the residue count of a given amino acid divided by the total sequence length) was then calculated for each metabolic enzyme. The pathway-level amino acid composition was defined as the arithmetic mean of enzyme-level composition vectors across all enzymes annotated to the pathway.

Calculation of amino acid enrichment scores for metabolic pathways

For each species, all metabolic enzymes were aggregated into a background metabolic enzyme pool. For each pathway k and amino acid a, we defined the

observed pathway-level statistic $T_{\mathrm{obs}}(k, a)$ as its pathway-level amino acid composition. We then performed a permutation test to quantify deviation from the background. Specifically, while keeping the number of enzymes in each pathway fixed, we repeatedly sampled the same number of enzymes from the background metabolic enzyme pool and computed the corresponding permuted statistic $T_{\mathrm{perm}}(k, a)$. This procedure was repeated 1,000 times to simulate a null distribution. The empirical one-sided p value was then defined as follows:

$$p(k, a) = Pr(T_{\mathrm{obs}}(k, a) \geq T_{\mathrm{perm}}(k, a)),$$

estimated by the fraction of permutations satisfying $T_{\mathrm{obs}}(k, a) \geq T_{\mathrm{perm}}(k, a)$. Under this definition, large p values ($p > 0.95$) indicate enrichment (higher-than-expected amino acid fraction), whereas small p values ($p < 0.05$) indicate depletion (lower-than-expected fraction) relative to the background pool.

To represent enrichment/depletion patterns on a unified scale, we transformed p values into an AA enrichment score: p values within the non-significant interval ($0.05 \leq p \leq 0.95$) were set to 0.5 (neutral), whereas tail p values ($p > 0.95$ or $p < 0.05$) were kept unchanged. This yielded an AA enrichment score matrix (rows: pathways; columns: the 20 standard amino acids).

Identification of AArchetypes

We performed hierarchical clustering of pathways based on Euclidean distance using the AA enrichment score matrix to identify candidate AArchetypes. For human, the metabolic pathways were clustered into four clusters. For other species, the number of clusters was determined by comparing distributions of silhouette coefficients across different numbers of clusters and selecting the number yielding highest average silhouette coefficient. KEGG functional category enrichment was performed for each

module in each species using one-sided Fisher's exact test to assess whether the module is enriched with major categories of metabolic functions such as carbohydrate metabolism, amino acid metabolism, and so on. Three indicators were considered in combination to identify AArchetypes in each species:

(1) Average silhouette coefficient, measuring the quality and robustness of clustering based on the AA enrichment score matrix.
(2) −log10(max(min enrichment p-value)), where for each module we took the minimum enrichment p-value across KEGG functional categories, then took the maximum across modules, and then computed the negative base-10 logarithm of this maximum. This metric assesses whether each module is enriched with at least one functional category.
(3) Ratio of significant part, defined as the fraction of non-neutral AA enrichment scores (i.e. $p(k,a)<0.05$ or $p(k,a)>0.95$) within the AA enrichment score matrix.

Species with an average silhouette coefficient greater than 0.15, a max(min enrichment p-value) below 0.05, and a ratio of significant part greater than 0.3 were considered to exhibit the emergence of AArchetypes.

Random forest classification of human enzymes across AArchetypes

To test whether the four human AArchetypes exhibit separable amino acid composition profiles at the protein level, we trained a random forest classifier using the amino acid composition of human metabolic enzymes as features and AArchetype assignments as labels. Classification performance was evaluated in a one-vs-rest setting for each AArchetype, in which each AArchetype was considered in turn as the positive class, while proteins from the remaining three AArchetypes were pooled as the negative class. Discriminative ability was quantified by the area under the receiver operating characteristic curve (AUC).

Correlation between amino acid composition and structure of human metabolic enzymes

To quantify the relationship between amino acid composition and 3D structure of human metabolic enzymes, we downloaded predicted structures (PDB format) from the AlphaFold database[13]. We then used Foldseek[14] to convert each enzyme structure into a length-matched per-residue structural sequence, represented by a 20-state structural alphabet that encodes local structural environments along the polypeptide chain. Pairwise Euclidean distances between these structural sequences were computed using DECIPHER[15], generating a structural distance matrix. We then computed pairwise Euclidean distances between amino acid composition vectors of the enzymes to obtain an amino acid composition distance matrix. Pearson correlation between these two distance matrices was then calculated over all off-diagonal elements.

Acquisition of proteome, amino acid exchange, and evolutionary datasets

Subcellular localization annotations for human metabolic enzymes and amino acid sequences of all proteins in *H. sapiens* and *S. cerevisiae* were downloaded from UniProt. For each species, we computed the proteome-level abundance of the 20 standard amino acids, defined as the average amino acid composition across all proteins in the proteome. Human proteins were dichotomized into membrane-associated proteins and non-membrane proteins based on UniProt membrane-related annotations. Protein abundance data in human tissue samples, including breast, skin, liver, lung, and kidney, were obtained from ProteomicsDB[16] (https://www.proteomicsdb.org/). Protein abundance datasets in human cell lines were obtained from CCLE[17] (https://sites.broadinstitute.org/ccle/). We computed the average abundance of each protein across tissues to get its mean expression level. Protein

abundance data for *S. cerevisiae* were obtained from a previous study by Brandon et al[18]. We used the mean protein abundance (molecules per cell) in the dataset as the expression level of yeast protein.

To quantify tissue specific expression, we calculated the Gini coefficient of tissue resolved protein abundance for each enzyme (higher Gini indicates stronger tissue specificity). Given the expression levels of a protein across n tissues $x_1, x_2, \ldots, x_n$, its mean expression level μ is:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Gini coefficient can be defined as the normalized mean absolute difference between all pairs of values:

$$Gini = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|$$

Differences in protein abundance and Gini coefficients between AArchetypes were evaluated using two-sided Wilcoxon rank-sum tests.

Metabolite uptake and secretion datasets in NCI-60 cell lines were obtained from a previous study by Aaron et al[19] to quantify the exchange fluxes of amino acids in human cells. Gene age data were obtained from the GenTree database[20], and evolutionary rate data were retrieved from the review by Zhang et al[21].

Comparison of amino acid composition between human enzymes and dietary patterns

We obtained amino acid composition profiles for ten human dietary patterns from our previous study[22]. Because experimental quantification of amino acids in foods often cannot reliably distinguish glutamate from glutamine, or aspartate from asparagine, fractions of these pairs of amino acids in human enzymes were merged before being compared to human dietary patterns. To

compare dietary patterns with pathway-level amino acid composition of the human metabolic network, we concatenated the relative abundance vectors from the two sets and performed t-distributed stochastic neighbor embedding (t-SNE; perplexity = 30) for visualization. We further quantified compositional similarity by computing Euclidean distances between each enzyme assigned to a given AArchetype and each dietary pattern in the amino acid composition space. For each (AArchetype, dietary pattern) pair, we summarized the resulting distance distribution across enzymes and used the median as the representative distance.

Calculation of protein cost and energy cost for amino acids

We quantified the protein and energetic costs required for the biosynthesis of non-essential amino acids under a physiologically relevant, high-growth condition. Specifically, we considered a nutrient-replete environment in which water, oxygen, inorganic salts, glucose, amino acids, and other essential nutrients (Table S1) are available. First, we used GECKO[23] to augment a human genome-scale metabolic model (Recon3D) with enzyme-usage constraints for individual reactions. For each metabolic enzyme, we computed the ratio of its molecular weight to its catalytic constant (MW/k_{cat}) to provide a proxy for the steady-state enzyme mass required to sustain one unit of metabolic flux through reactions catalyzed by it. For each metabolic reaction, a reaction-level protein cost coefficient $c_{p,r}$ was computed by the minimal MW/k_{cat} among all isozymes catalyzing this reaction. This reaction-level protein cost coefficient quantifies the minimal mass of enzyme required to carry one unit of flux through this reaction. We also defined a reaction-level energy cost coefficient $c_{e,r}$ of each reaction, quantified by the number of ATP molecules hydrolyzed in this reaction. Let v_r denote the flux through reaction r, the protein and energy costs consumed by this reaction can be calculated by $c_{p,r}v_r$ and $c_{e,r}v_r$, respectively. Collecting all reaction-level coefficients into

vectors $\boldsymbol{c}_p$ and $\boldsymbol{c}_e$, the total protein cost and energy cost corresponding to a metabolic flux configuration $\boldsymbol{v}$ are $\boldsymbol{c}_p^T\boldsymbol{v}$ and $\boldsymbol{c}_e^T\boldsymbol{v}$, respectively.

We next applied flux balance analysis (FBA) to compute the marginal protein and energy costs associated with the biosynthesis of each non-essential amino acid. We formulated the problem as solving an optimal flux configuration $\boldsymbol{v}$ minimizing the total protein cost $\boldsymbol{c}_p^T\boldsymbol{v}$ or energy cost $\boldsymbol{c}_e^T\boldsymbol{v}$ under flux balance, sustained growth, and other constraints. The following constraints were included in the optimization:

$$\begin{cases} \boldsymbol{S}\boldsymbol{v} = \boldsymbol{0} \\ \boldsymbol{l} \leq \boldsymbol{v} \leq \boldsymbol{u} \\ v_{BM} \geq b \end{cases}$$

In which $\boldsymbol{S}$ is the stoichiometric matrix, $\boldsymbol{l}$ and $\boldsymbol{u}$ are lower and upper bounds of fluxes, v_{BM} is the biomass synthesis flux, b is a preset positive lower bound of the biomass synthesis flux. Default bounds of reactions were set to [0,1000] unless specified later. Exchange reactions (i.e. reactions with the form "⇒X[e]", representing importing a metabolite from the external environment into the extracellular space) for molecules other than the growth-supporting nutrients listed in Table S1 were disabled by setting both their lower and upper bounds to 0. Upper bounds for exchange reactions for growth-supporting nutrients in Table S1 were set by referencing nutrient intake under human dietary patterns[22].

We first computed the maximal feasible biomass synthesis flux $v_{BM,\max}$ under these constraints by solving the linear programming problem below:

$$\text{Maximize } v_{BM}$$
$$s.t. \begin{cases} \boldsymbol{S}\boldsymbol{v} = \boldsymbol{0} \\ \boldsymbol{l} \leq \boldsymbol{v} \leq \boldsymbol{u} \end{cases}$$

The lower bound of biomass synthesis flux b was then set to $0.5v_{BM,\max}$, thereby constraining growth to at least 50% of the maximal possible rate.

For each non-essential amino acid A_i, upper bound of its uptake flux was set to 0 to enforce its *de novo* synthesis. We first computed the protein cost P_0 and energy cost E_0 at a reference state under the high-growth constraint:

$$\text{Minimize } \boldsymbol{c}_p^T\boldsymbol{v} \text{ or } \boldsymbol{c}_e^T\boldsymbol{v}$$
$$s.t.\begin{cases} \boldsymbol{S}\boldsymbol{v} = \boldsymbol{0} \\ \boldsymbol{l} \le \boldsymbol{v} \le \boldsymbol{u} \\ v_{BM} \ge 0.5 v_{BM,\max} \end{cases}$$

Next, to quantify the marginal cost of A_i, a small positive number δ was added to its stoichiometric coefficient α_i in the biomass reaction to simulate an A_i-demanding state. We then recomputed the minimal protein and energy costs $P_{1(i)}$ and $E_{1(i)}$ at this A_i-demanding state by minimizing $\boldsymbol{c}_p^T\boldsymbol{v}$ and $\boldsymbol{c}_e^T\boldsymbol{v}$. Finally, marginal protein and energy costs of the amino acid A_i were calculated as follows:

$$P_{A_i} = \frac{P_{1(i)} - P_0}{\delta}$$
$$E_{A_i} = \frac{E_{1(i)} - E_0}{\delta}$$

Cost analysis of metabolic enzymes and pathways

For a metabolic enzyme with amino acid composition vector $\boldsymbol{f}$, in which f_i means the fraction of amino acid A_i in the sequence of that enzyme, its protein and energy costs were computed as follows:

$$P = \sum_{i=1}^{20} P_{A_i} f_i$$
$$E = \sum_{i=1}^{20} E_{A_i} f_i$$

To evaluate how small changes in fraction of an individual amino acid affect the average protein cost of a metabolic pathway, we performed a cost sensitivity analysis for each pair of an amino acid and a metabolic pathway. Let $\boldsymbol{f}$ denote the average amino acid composition of enzymes in that pathway, $\boldsymbol{P}_A$ and $\boldsymbol{E}_A$ denote the vectors of protein and energy costs of amino acids, the

protein and energy costs of this pathway were calculated by $\boldsymbol{P}_A^T\boldsymbol{f}$ and $\boldsymbol{E}_A^T\boldsymbol{f}$, respectively. Next, to compute the protein and energy cost sensitivity score for amino acid A_i, we randomly generated n amino acid compositions $\boldsymbol{f}_1'$, $\boldsymbol{f}_2'$, …, $\boldsymbol{f}_n'$, where the fraction of A_i was set to $f_i + \delta$, δ is a small positive number. The cost sensitivity scores of A_i for this pathway ($s_{P,i}$ for protein cost, $s_{E,i}$ for energy cost) were then computed as follows.

$$s_{P,i} = \frac{1}{\delta}\left(\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{P}_A^T\boldsymbol{f}_k' - \boldsymbol{P}_A^T\boldsymbol{f}\right)$$

$$s_{E,i} = \frac{1}{\delta}\left(\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{E}_A^T\boldsymbol{f}_k' - \boldsymbol{E}_A^T\boldsymbol{f}\right)$$

IDR and LLPS-associated analysis

Intrinsically disordered region (IDR) annotations for human proteins were obtained from the literature[24]. IDR proportion of a protein was calculated by dividing the length of IDR within it by its total length. LLPS protein annotations for multiple species were retrieved from the DrLLPS database[25]. The four AArchetypes were further grouped into two pairs based on whether they are dominated by membrane proteins: AArchetypes A and B (non-membrane proteins), and AArchetypes C and D (membrane proteins). For each pair, we counted the numbers of LLPS and non-LLPS proteins and performed hypergeometric tests to assess LLPS protein enrichment for each AArchetype. To test whether IDR-containing proteins (IDR proteins) and phase-separating proteins (LLPS proteins) are associated with higher protein cost at the proteome level, we carried out gene set enrichment analysis (GSEA) for the two gene sets. The human proteome was ranked by protein cost from high to low, and this ranked gene list was used as input for GSEA against KEGG pathways as well as the two gene sets corresponding to IDR proteins and LLPS proteins. The resulting normalized enrichment scores (NES) and visualization of their distributions were used to assess the enrichment.

Mathematical model for internalization of LLPS externality

The model involves two groups of amino acids: an LLPS-related amino acid group X with average cost of w, and an LLPS-unrelated amino acid group Y with average cost of 1, that together forms a protein with abundance A. Therefore, the total cost of the protein can be written as:

$$C(a, A) = \big(wa + (1 - a)\big)A$$

In which a is the fraction of the LLPS-associated amino acid group X in the protein. When the total biosynthetic cost of amino acids used in producing this protein is constrained by a budget (set as 1 for simplicity), we have an inequality constraint as follows:

$$\big(wa + (1 - a)\big)A \leq 1$$

Next, we estimated the probability for this protein to undergo LLPS as a function $p(a)$ by fitting a logistic function using sequences and LLPS capabilities of human proteins[25] (see "Logistic regression model for LLPS probability" for details):

$$p(a) = \frac{1}{1 + e^{6.3 - 10.6a}}$$

We therefore define a functional payoff function $P_{\text{func}}(a, A)$ as the sum of the baseline benefit associated with the protein abundance A and the marginal increase of reward caused by the LLPS portion:

$$P_{\text{func}}(a, A) = A(dp(a) + 1)$$

In which d, termed the LLPS payoff gain, quantifies the increase in the payoff function contributed by one additional protein undergoing LLPS. This functional payoff function quantifies how much the function of the protein is enhanced by LLPS. The overall payoff function of the protein is then defined as the weighted sum of $P_{\text{func}}(a, A)$ and an information capacity term $P_{\text{info}}(a)$ that quantifies the ability of the protein sequence to store information:

$$P(a, A) = P_{\text{func}}(a, A) + \beta P_{\text{info}}(a)$$

We assume that the protein determines its amino acid composition by maximizing the payoff function under the constraint of limited budget:

$$\text{Maximize } P(a, A) = P_{\text{func}}(a, A) + \beta P_{\text{info}}(a)$$
$$s.t. \begin{cases} \big(wa + (1-a)\big)A \leq 1 \\ 0 \leq a \leq 1 \end{cases}$$

As $P_{\text{func}}(a, A)$ is a linear function of the protein abundance A with the positive linear factor $dp(a) + 1$, the payoff function $P(a, A)$ is also a monotonically increasing function of A. Therefore, at the optimal solution, A equals to its maximally allowed value under the limited budget constraint:

$$A_{\max}(a) = \frac{1}{wa + (1-a)}$$

Therefore, the optimal payoff function can be computed by optimizing a:

$$\text{Maximize } P(a) = P_{\text{func}}\big(a, A_{\max}(a)\big) + \beta P_{\text{info}}(a)$$
$$s.t.\ 0 \leq a \leq 1$$

The constrained optimization problem was solved by the MATLAB built-in function “fmincon”.

Logistic regression model for LLPS probability

LLPS labels for human proteins were defined based on information in the DrLLPS database (1 for LLPS protein and 0 otherwise). We partitioned the 20 amino acids into two sets according to their association with IDR and LLPS: X (LLPS-related amino acids: K, A, Q, P, G, E, S) and Y (all remaining amino acids). For the i-th protein with amino acid composition f_i, its fraction of LLPS-related amino acids was calculated as follows:

$$a_i = \sum_{k \in X} f_{i,k}$$

To model how the probability of phase separation depends on a, we fitted a binary logistic regression model with a as the predictor and the LLPS label y as the outcome:

$$p(a) = \frac{1}{1 + \exp(6.3 - 10.6a)}$$

Definition of information capacity term

The information capacity term of a protein was defined as a function $P_{\text{info}}(a)$ of fraction of LLPS-related amino acids in the sequence of this protein, a. To derive the analytical expression of $P_{\text{info}}(a)$, we started from computing the Shannon entropy of a protein from its composition of 20 standard amino acids, $\boldsymbol{f} = (f_1, f_2, \dots, f_{20})$:

$$H(\boldsymbol{f}) = -\sum_{i=1}^{20} f_i \log_2 f_i$$

For a fixed sequence length, the logarithm of the number of sequences compatible with a given composition is proportional to $H(\boldsymbol{f})$. The information capacity $P_{\text{info}}(a)$ was then defined as maximal $H(\boldsymbol{f})$ among all possible sequences satisfying $\sum_{i \in X} f_i = a$ (i.e. total fraction of amino acids in the LLPS-related group X equals to a):

$$P_{\text{info}}(a) = \max_{\sum_{i \in X} f_i = a} -\sum_{i=1}^{20} f_i \log_2 f_i$$

As group X contains 7 amino acids, the sequence achieves maximal Shannon entropy when the fraction a is uniformly distributed across the 7 amino acids in X (LLPS-related amino acids) and the remaining fraction $1 - a$ is uniformly distributed across the 13 amino acids in Y (LLPS-unrelated amino acids). Therefore, the information capacity corresponding to a can be calculated:

$$\begin{aligned} P_{\text{info}}(a) &= 7\left(-\frac{a}{7}\log_2\frac{a}{7}\right) + 13\left(-\frac{1-a}{13}\log_2\frac{1-a}{13}\right) \\ &= -a\log_2\frac{a}{7} - (1-a)\log_2\frac{1-a}{13} \end{aligned}$$

This function therefore quantifies how the size of the accessible sequence space varies as the fraction of LLPS-related residues changes.

Estimation of species-specific LLPS cost ratio

For *H. sapiens* and *S. cerevisiae*, we first computed their average amino acid composition over the entire proteome. Let X and Y denote the groups of LLPS-related and LLPS-unrelated amino acids, the LLPS cost ratio w was defined as follows:

$$w = \frac{\sum_{i \in X} P_{A_i} f_i}{\sum_{i \in Y} P_{A_i} f_i}$$

In which P_{A_i} is the marginal protein cost of amino acid A_i in this species (human or yeast), f_i is the average fraction of amino acid A_i in proteins from this species.

Inference of parameters $\boldsymbol{\beta}$ and $\boldsymbol{d}$

The model parameter β was inferred based on the assumption that in yeast ($w = 0.26$) where LLPS proteins are only a small fraction, a protein with median a ($a = 0.41$) does not benefit from LLPS. This assumption means that under the parameter set $\{w = 0.26, d = 0\}$, the optimal solution of a is 0.41. Hence, we sought to determine the parameter β that gives this optimal solution. We computed optimal values of a (referred to as $a_{\mathrm{opt}}(\beta)$) for an array of 1001 evenly distributed β values in the range of [0,10], $\{\beta_i = 0.01(i-1)\}$, and then selected the two values β_i and β_{i+1} that satisfies $a_{\mathrm{opt}}(\beta_i) > 0.41$ and $a_{\mathrm{opt}}(\beta_{i+1}) < 0.41$. Assuming that a_{opt} changes linearly with β in the range $[\beta_i, \beta_{i+1}]$, the parameter β corresponding to $a_{opt} = 0.41$ was computed as follows by inverse linear interpolation:

$$\beta(a_{opt} = 0.41) = \beta_i + (\beta_{i+1} - \beta_i)\frac{0.41 - a_{\mathrm{opt}}(\beta_i)}{a_{\mathrm{opt}}(\beta_{i+1}) - a_{\mathrm{opt}}(\beta_i)}$$

Similarly, to infer the parameter d for a human protein with LLPS-related amino acids fraction a, optimal values of a (referred to as $a_{\mathrm{opt}}(d)$) were computed for 3001 evenly distributed values of d in the range of [0,30]. For

each given a in the range $[a_{\mathrm{opt}}(0), a_{\mathrm{opt}}(30)]$, its corresponding value of d was also computed using inverse linear interpolation.

Results

Emergence of AArchetypes in human metabolic enzymes

To test whether the amino acid composition of human proteins still follows a single cost-minimization principle, we started from metabolic enzymes that constitute a large fraction of the human proteome[26]. We obtained amino acid sequences of all human metabolic enzymes from the KEGG database and quantified pathway-level amino acid composition by computing an amino acid (AA) enrichment score using permutation test (Fig. 1a, Methods). Hierarchical clustering of the amino acid enrichment scores of KEGG metabolic pathways identified four discrete clusters each characterized by a distinct pattern of amino acid usage (Fig. 1b). Significant enrichment of major metabolic categories was observed in each cluster: cluster A is enriched with carbohydrate metabolism, cluster B is enriched with amino acid metabolism, cluster C is enriched with xenobiotics metabolism, and cluster D is enriched with lipid and glycan metabolism (Fig. 1c). Furthermore, although the analysis was performed at the level of metabolic pathways, individual metabolic enzymes in each cluster can be distinguished from those in remaining clusters solely based on their amino acid composition (Fig. 1d), confirming that the diversification of amino acid composition occurs at the level of single proteins rather than being an artifact of pathway-level averaging.

To test whether these clusters simply reflect structural similarity between functionally related proteins, we computed pairwise distances in amino acid composition with structural distances between metabolic enzymes (Fig. 1e, Methods). These two distances were poorly correlated (Pearson correlation =

-0.047, Fig. 1e), suggesting that structural similarity does not explain the observed clustering of amino acid composition.

Together, these results reveal that human metabolic enzymes exhibit the emergence of four distinct amino acid archetypes (AArchetypes), each associated with specific metabolic functions independent of protein structure. This indicates existence of forces beyond catalytic requirements or structural constraints that shape amino acid composition of human metabolic enzymes, pointing to underlying economic principles that act differently across metabolic sectors.

AArchetypes only emerge in multicellular organisms

Next, we sought to determine whether the emergence of AArchetypes is a universal phenomenon conserved across the tree of life or a feature of complex life. Hence, we analyzed metabolic enzymes from ten phylogenetically diverse species spanning prokaryotes and eukaryotes, unicellular and multicellular organisms, plants and animals (Fig. 2a). For each species, we computed pathway-level average amino acid compositions and performed classical multidimensional scaling (MDS) to visualize the global landscape of amino acid composition across species (Fig. 2b). Multicellular organisms, such as human and mouse, exhibit similarity in pathway-level amino acid composition, while unicellular organisms such as bacteria and yeast display heterogeneous distributions (Fig. 2b). This difference suggests that the diversification into discrete AArchetypes that we observed in human might be a feature of multicellular life.

To test this hypothesis quantitatively, we applied the same enrichment and clustering framework used for human enzymes to all species (Fig. 2c,d, S1a-g). We used three independent metrics to assess the presence of AArchetypes: (1) average silhouette coefficient, which quantifies the tendency

of forming separate clusters; (2) maximal minimum (max-min) of enrichment p-values, which evaluates whether all clusters are enriched with metabolic categories; and (3) fraction of significant entries in the amino acid enrichment scores, which quantifies the heterogeneity of amino acid usage across metabolic pathways. These metrics together assess whether metabolic enzymes in a species segregate into well-separated and functionally meaningful AArchetypes.

Using this framework, we found that AArchetypes emerged exclusively in multicellular organisms (Fig. 2c,d, S1h), including human (*Homo sapiens*, hsa), chimpanzee (*Pan troglodytes*, ptr), mouse (*Mus musculus*, mmu), fruit fly (*Drosophila melanogaster*, dme), worm (*Caenorhabditis elegans*, cel), and plant (*Arabidopsis thaliana*, ath), but were absent within all unicellular organisms examined. The robust emergence of AArchetypes in multicellular organisms spanning a diversity of phylogenetic branches implies that diversification of amino acid composition in metabolic enzymes is driven by evolutionary forces that occur only when organisms transition from single cells to differentiated multicellular systems.

Youngest AArchetype minimizes synthetic cost

To characterize how amino acid economics differs across human AArchetypes, we computed the synthetic costs of all non-essential amino acids within the human metabolic network. Two types of costs were considered: protein cost, which quantifies the amount of enzymes needed for *de novo* synthesis of a unit of the amino acid, and energy cost, which quantifies the number of ATP molecules consumed in synthesizing a unit of the amino acid. Both costs were defined as marginal increases in total enzyme usage or ATP consumption upon a small increase in demand for each amino acid (Fig. 3a, S2a, Methods).

The two costs are uncorrelated at single amino acid level (Pearson correlation = 0.006, Fig. 3b), indicating that amino acids expensive in terms of enzymatic investment are not necessarily energy intensive, and *vice versa*. To validate these theoretical costs, we examined the uptake and secretion profiles of amino acids in the NCI-60 human cell lines. Amino acids with highest protein and energy costs, serine (highest protein cost) and arginine (highest energy cost), are also among those most actively taken up from the environment (Fig. 3c). This observation suggests an intuitive strategy for the cell to bypass high synthetic costs of amino acids through import. Applying the same framework to the budding yeast *S. cerevisiae* reproduced the known negative correlation[4] between protein cost and proteome abundance of amino acids (Fig. 3d, top), therefore confirming the validity of our approach. However, although increase in synthetic cost is weakly correlated to decrease in proteome abundance in human compared to yeast (Fig. S2b), protein cost and proteome abundance of amino acids correlated poorly in human (Fig. 3d, bottom), suggesting that cost minimization alone no longer governs amino acid usage in the human proteome.

We next asked whether cost minimization might operate selectively within specific AArchetypes. To our surprise, although protein and energy costs correlated poorly at single amino acid level (Fig. 3b), the four AArchetypes exhibit consistent ordering when these two costs were aggregated at the AArchetype level: AArchetype C is the most cost-effective, followed by AArchetype A, while AArchetypes B and D are the costliest (Fig. 3e). Therefore, the convergence of protein and energy costs at the level of AArchetypes highlights intrinsic economic strategies that emerge only at the AArchetype level.

To understand the divergence in cost across AArchetypes, we estimated the age and evolutionary rate of human metabolic enzymes (Methods) and found

that the most cost-effective AArchetype C was enriched with younger proteins (Fig. 3f) that evolve faster (Fig. 3g). This pattern suggests that in human metabolic enzymes, the strategy of cost-saving is largely constrained to a small subset of youngest proteins. Ancient and highly conserved enzymes that originated before the appearance of complex multicellular organisms might have already adapted their amino acid composition to the ancient metabolic network of unicellular ancestors, therefore unable to freely reoptimize towards minimal synthetic cost under the modern human metabolic network.

Central carbon metabolism AArchetype matches dietary amino acid composition

Dietary intake is an important source of amino acids in animals including humans. Amino acids obtained from diet can bypass the process of *de novo* synthesis, therefore reducing the associated energy and protein costs. This raises the possibility that, beyond cost-saving through biosynthetic optimization, some enzymes may achieve an effectively “cheaper” amino acid supply by matching their amino acid composition to that of dietary patterns — a strategy analogous to trading in economics. To test this hypothesis, we compared the average amino acid compositions of metabolic pathways in each AArchetype to amino acid compositions of ten representative human dietary patterns[22], including both dietary patterns associated beneficial health effects (e.g. Mediterranean and Japanese diets) and those associated with potential health risks (e.g. American diet).

Dimensionality reduction of amino acid composition of human dietary patterns together with human pathway-level amino acid compositions using t-distributed neighbor embedding (t-SNE, Fig. 4a) indicates that amino acid composition of AArchetype A, which is enriched for carbohydrate and central carbon metabolism, was closest to that of human dietary patterns (Fig. 4a).

This qualitative observation suggests a preferential alignment between amino acid composition of human diet and enzymes participating in central carbon metabolism. Next, we computed Euclidean distances in amino acid composition between each dietary pattern and metabolic enzymes in each AArchetype, and summarizing these distances by their median (Fig. 4b). Across all ten dietary patterns, AArchetype A consistently exhibited the smallest median distance to dietary amino acid composition. The robustness across diverse dietary patterns reflects a broad compositional compatibility between dietary amino acid supply and the amino acid composition of central carbon metabolism enzymes. Notably, dietary patterns with balanced consumption of a diverse range of minimally processed foods, such as Mediterranean, Japanese, and Paleo diets, also show amino acid composition closer to that of all four AArchetypes (Fig. 4b). This is consistent with previous findings in model organisms such as fruit fly that proteome-matching dietary amino acid compositions have beneficial health outcomes[9].

Together, these results support a “trading” strategy underlying AArchetype A: instead of minimizing biosynthetic costs through sequence-level optimization (as observed for the youngest and fastest-evolving AArchetype C), central carbon metabolism enzymes, which have originated in the ancient metabolic networks and conserved in evolution and express in high abundance, may reduce their effective amino acid costs by relying on dietary supply of amino acids that already approximates their compositional requirements.

Costly AArchetypes are associated with structural disorder and phase separation

We next asked whether the high costs of AArchetypes B and D reflect additional biophysical or cellular constraints that require incorporation of high-cost amino acids. For instance, transmembrane domains of membrane proteins require hydrophobic amino acids, therefore shaping their amino acid

composition[27, 28]. Membrane association partly explains the emergence of AArchetypes: AArchetypes C and D were strongly enriched for membrane proteins compared to AArchetypes A and B (Fig. 5a). However, the two high-cost AArchetypes involve both soluble (B) and membrane-associated (D) enzymes (Fig. 5a), indicating that costly enzymes exist within both localization contexts.

Using quantitative proteomics data across human tissues (Methods), we found that enzymes within the two high-cost AArchetypes are expressed at lower abundance relative to their counterparts within the same localization group: B is lower than A among non-membrane enzymes, and D is lower than C among membrane-enriched enzymes (Fig. 5b). Moreover, protein abundances of enzymes in the costly AArchetypes B and D are tissue specific as reflected by higher Gini coefficients (Fig. 5c, Methods).

Together, these results suggest that enzymes with costly amino acid composition tend to reduce their realized cost by lowering overall abundance and restricting expression to specific tissues. However, such association alone does not explain the existence of costly AArchetypes, as adopting a less costly amino acid composition can always reduce the cost of enzymes regardless of the protein abundance. To identify features associated with these costly AArchetypes, we next sought to determine single amino acids contributing to increased cost of enzymes. By computing protein and energy cost sensitivity scores quantifying the change in pathway-level costs with a marginal compositional increase of each single amino acid (Fig. 5d, S3a, Methods), we identified a subset of amino acids associated with the costly AArchetypes, including proline, glycine, glutamate, glutamine, arginine, asparagine, and serine (Fig. 5d). This set overlaps substantially (5 out of 7) with amino acids that are frequently enriched in intrinsically disordered regions (IDRs), including lysine, alanine, glutamine, proline, glycine,

glutamate, and serine[29] (Fig. 5d), establishing a connection between high synthetic cost and intrinsic structural disorder.

Conformational flexibility of IDRs is known to support multivalent, weak interactions with proteins and nucleic acids, which is a key molecular basis for the formation of dynamic biomolecular condensates by liquid-liquid phase separation (LLPS)[30]. Such condensates can carry functional benefits for metabolic enzymes by concentrating reactants, tuning reaction kinetics, and enabling dynamic, reversible compartmental organization[31-34]. Nevertheless, they also carry risk: excessive or aberrant condensation can promote protein misfolding and aggregation and has been implicated in pathologies including neurodegeneration and cancer[35-37]. Therefore, we hypothesize that the high cost of IDR and LLPS-associated amino acids might help suppress the potential accumulation of macromolecular condensates formed by LLPS-prone proteins. We found that for the membrane-unrelated AArchetypes A and B, enzymes in the higher-cost AArchetype B exhibit a higher proportion of IDR (Fig. 5e) and fraction of LLPS proteins (Fig. 5f) compared to those in the lower-cost AArchetype A. Such association between higher cost and preference of IDR and LLPS was also observed while comparing the membrane-associated AArchetypes C and D (Fig. 5e,f), indicating a connection between structural disorder and high cost independent of localization context of enzymes.

To test whether this coupling extends beyond metabolic enzymes, we performed GSEA to identify functional categories associated with high cost throughout all human proteins. Compared to all KEGG pathways, the two sets of IDR and LLPS proteins ranked the top in GSEA enrichment scores (Fig. 5g, S3b-e), indicating that involvement of IDR and capability of LLPS are the most pronounced features of high-cost proteins. Direct proteome-wide comparisons

further confirmed that LLPS proteins have significantly higher protein and energy cost than non-LLPS proteins (Fig. 5h).

Taken together, the association between high synthetic cost and LLPS suggests an economic interpretation for why costly AArchetypes B and D persist. LLPS can be functionally valuable for enzymes with this capability but potentially hazardous at the entire cell level, therefore creating a negative externality. A plausible way for the cell to internalize this externality is to effectively raise the costs of LLPS-associated amino acids, thereby making LLPS-prone proteins intrinsically expensive to produce and favor regulatory strategies that limit their total burden. In this framework, AArchetypes B/D complement cost saving (C) and trading (A) as a third strategy: internalization of externality, where amino acids associated with negative externality are subject to enhanced synthetic costs.

A mathematical model for LLPS externality internalization by increasing synthetic cost

The coupling between high synthetic cost and IDR/LLPS propensity in the human AArchetypes motivates an economic interpretation: LLPS can provide protein-level functional benefits, but excessive or dysregulated condensation could impose a burden at the cellular level. In economic theory, such a mismatch between local benefit and global burden is described by negative externality, which can be internalized through price-based instruments such as Pigouvian taxation. We therefore constructed a simplified mathematical model to elucidate how elevated synthetic costs of LLPS-related amino acids can act as an effective pricing mechanism that restrains LLPS level while preserving functional gains (Fig. 6a, Methods). We used the model to examine how amino acid prices shape the optimal amino acid composition and protein abundance. We also performed a counterfactual simulation in

which human proteins optimize their amino acid compositions under a pricing system that does not elevate the costs of LLPS-related amino acids.

The model involves two groups of amino acids: an LLPS-related amino acid group X and an LLPS-unrelated amino acid group Y, that together forms a protein with abundance A. The protein determines its fraction of LLPS-related amino acids, a, by maximizing a payoff function under the constraint of limited budget. LLPS probability increases with a, generating an LLPS-dependent functional gain. The payoff function involves a functional benefit term related to the protein abundance A and LLPS-dependent functional gain (Fig. S4a, Methods), and an information capacity term quantifying the sequence complexity:

$$\text{Maximize } P(a, A) = P_{\text{func}}(a, A) + \beta P_{\text{info}}(a)$$
$$s.t.\ \big(wa + (1 - a)\big)A \leq 1$$

The model contains two LLPS-related parameters: the LLPS payoff gain d and the LLPS cost ratio w. The LLPS payoff gain d is a protein-specific parameter quantifying the increase in the payoff function contributed by one additional protein undergoing LLPS. The LLPS cost ratio w is a species-specific parameter estimated by comparing synthetic costs of LLPS-related and LLPS-unrelated amino acids (Fig. 6a, Methods). A value of w over 1 indicates a higher cost of LLPS-related amino acids. LLPS cost ratio w differs by one order of magnitude between human (*H. sapiens*) and yeast (*S. cerevisiae*): w=5.77 for human and 0.26 for yeast (Fig. 6b), indicating higher costs of LLPS-related amino acids in human but not in yeast. As the yeast proteome contains a low fraction (around 0.1, Fig. S4b) of LLPS proteins, we assumed that a yeast protein with median a ($a = 0.41$) has zero LLPS payoff gain ($d = 0$) and estimated the parameter β accordingly (Fig. S4c, d, Methods).

The difference in LLPS cost ratio w between human and yeast results in distinct modes of amino acid compositional optimization. To illustrate such distinction, we computed the maximal payoff function under varying a, $P_{\max}(a)$, for both human (w=5.77, Fig. 6c) and yeast (w=0.26, Fig. 6d). For human, the payoff function first increases and then decreases with a independent of the LLPS payoff gain d, as the higher cost of LLPS-related amino acids in human strongly penalizes a close to 1 by reducing the capacity of A. On the other hand, for yeast the payoff function only peaks at $a < 1$ for smaller values of d and becomes monotonically increasing for higher d. This is because in the absence of the penalty by costly LLPS-related amino acids, the only factor that prevents a from becoming 1 is the low information capacity for a near 1 (Fig. S4e), which can be compensated by the protein abundance A (Fig. S4f) and LLPS reward $dp(a)$ (Fig. S4a) that both increase with a. Moreover, the model-predicted positive coupling between a and A in yeast was also validated by proteomics data (Fig. S4g, Methods), while such relationship diminishes in human (Fig. S4h), therefore further supporting our model.

Simulation of the relationship between the optimal solution of a, a_{opt} and the LLPS payoff gain d for human and yeast (Fig. 6e) shows that a_{opt} rapidly approaches 1 when d increases in the yeast model (w=0.26), while slowly grows from around 0.3 to around 0.7 in the human model (w=5.77), suggesting that in the yeast model where LLPS-related amino acids do not have elevated costs, a small functional reward by LLPS pushes a_{opt} towards 1. In contrast, for the human model, the higher costs of LLPS-related amino acids create a trade-off between increased LLPS reward and decreased protein abundance when a increases, allowing gradual tuning of a_{opt} under varying d. In such scenario, proteins that benefit more from LLPS (i.e.

proteins with higher d) will be enriched with LLPS-related amino acids, resulting in a right-skewed distribution of a.

Consistent with the prediction, distributions of a in human (Fig. 6f) and yeast (Fig. 6g) proteins show distinct patterns: the human distribution is right-skewed (skewness = 0.51) while the yeast distribution is not (skewness = -0.20). The range of a in real human proteins also aligns well with the model-predicted range of a_{opt}, further supporting the model and allowing direct inference of the LLPS payoff gain d for almost all human proteins (19748 out of 20403, Fig. 6h, Methods). We therefore inferred LLPS payoff gain d for those human proteins, and performed a counterfactual simulation in which human proteins retain their inferred LLPS payoff gains but are optimized under the yeast-like pricing regime. In this scenario, optimal a for almost all human proteins becomes near 1 (a>0.99 for 19482 out of 19748 proteins, Fig. 6i). This outcome increases protein-level payoff but would dramatically expand the burden of LLPS-related condensate formation at the cellular level, consistent with a scenario of negative externality. Taken together, the model provides a quantitative illustration of how the elevated synthetic costs of LLPS/IDR-related amino acids can internalize LLPS-related negative externalities.

Discussion

In this study, we reveal economic strategies underlying the diversification of amino acid composition of metabolic enzymes into AArchetypes, i.e. functionally related clusters of metabolic enzymes that share similar amino acid compositions. Catalyzing specific biochemical reactions, the primary goal of metabolic enzymes, does not require the amino acid composition to be strictly fixed[38]. Therefore, like how Carl Jung's archetypes reflect conserved, universal patterns in the psychological architecture of all human beings, the

emergence of AArchetypes in metabolic enzymes indicates previously unexplored, catalysis-independent forces that shape amino acid sequences of enzymes, such as evolutionary history, environmental conditions, and economic strategies in complex multicellular organisms.

Across ten phylogenetically distant species, robust AArchetype emergence was observed in animals and plants but not in the unicellular species. The observation suggests, unlike the principle of synthetic cost minimization in unicellular organisms such as bacteria and yeast, minimization of synthetic cost no longer governs the global amino acid composition of multicellular proteome. Using human metabolic enzymes as a model system for multicellularity, our analyses define two alternative economic strategies: trading, in which dietary amino acid composition matches that of conserved and abundant central carbon metabolic enzymes; and internalization of negative externality, in which LLPS-associated amino acids exhibit enhanced synthetic costs to prevent excessive condensate formation.

The emergence of these two economic strategies reflects fundamental differences between unicellular and multicellular organisms. Unicellular organisms and multicellular organisms differ intrinsically in strategy of nutrient acquisition. While unicellular organisms often rely on *de novo* biosynthesis from fluctuating, opportunistic supply of small molecule nutrients to meet their metabolic demands, nutrient supply in multicellular organisms becomes increasing buffered and internalized through organism-level systems for digestion, absorption, and long-range transport between tissues. Such systems exist across the kingdom of multicellular life, spanning digestion and circulation systems in animals, root, xylem, and phloem in plants, and interconnected mycelial networks in fungi[39]. This transition transforms amino acids from stochastic environmental inputs into regulated resources, enabling a shift from purely biosynthetic cost minimization toward a trading-based

strategy in which external supply can partially substitute for endogenous production.

Second, the enhanced synthetic costs of LLPS-related amino acids in human metabolism suggests a strategy of LLPS externality internalization. Negative externality, a concept borrowed from economics, describes the harmful impacts that actions of business firms have on others. An example is a factory that pollutes the air while making its products, as the factory benefits from selling the products while society suffers from the pollution. In economic theory, negative externalities are commonly internalized through three strategies: price-based instruments, quantity control, and regulatory mechanisms. Price-based instruments, such as Pigovian tax, increase the private cost associated with activities generating social harm[40]. Quantity-based instruments restrict total allowed scale of such activities, while regulatory mechanisms provide avenues for elimination of the harmful effects, limiting long-term accumulation and systemic risk. Both quantity-based and regulatory strategies have clear counterparts in the cellular control of biomolecular condensates. Dosage-sensitive genes tend to encode LLPS proteins and IDR-containing proteins[41-43], suggesting that the total allowed amount of LLPS proteins is tightly controlled by selective pressure disfavoring their amplification; IDR-containing proteins have shorter half-lives[44], limiting their steady state concentration. These observations together point to abundance limitation as a quantity-control mechanism for managing LLPS-associated risks. Multiple regulatory mechanisms co-exist to facilitate the disassembly and clearance of condensates[45], including post-translational modification[46, 47], autophagy[48], and molecular chaperones[49]. These mechanisms, together with the Pigovian taxation of LLPS-related amino acids in our theory, form a multi-layer economic system internalizing the negative externality associated with LLPS.

Even when a well-folded catalytic domain is sufficient for catalysis, disordered regions and phase separation can still provide an additional layer of control over its dynamics, location, and specificity. Regarding the enhancement of enzymatic activity by LLPS, current studies suggest that this effect is not universal, but highly conditional on the specific reaction system, substrate properties, scaffold organization, and the local physicochemical environment within condensates; under different conditions, LLPS can increase reaction rates, inhibit activity, or alter reaction specificity[33, 34, 50]. Our mathematical model is explicitly designed to capture this context-dependent functional reward of LLPS. It represents the potential functional benefit of LLPS with a protein-specific parameter d that serves as a simplified abstraction of a context-dependent LLPS reward, rather than assuming that LLPS uniformly and inevitably enhances enzyme activity.

Among the LLPS-related amino acids subject to Pigovian taxation in human metabolism, serine is the one with highest protein cost predicted by our theory. This is consistent with the biophysical role of serine in affecting material properties of molecular condensates. Serine residues promote hardening of condensates, generating gel-like or solid-like states[51, 52]. Such liquid-to-solid transition can cause abnormal accumulation of condensates associated with a variety of neurodegenerative diseases[37, 53]. Such association between serine and severest negative externality of LLPS explains why serine is subject to strongest Pigovian taxation under our theoretical framework.

It is worth clarifying that although IDRs often participate in the process of LLPS, these two are not equivalent. Long before the emergence of the LLPS concept, IDRs had already been recognized as functional structural elements capable of mediating flexible linkers, multivalent and reversible interactions, and dense post-translational regulation[54-56]. Therefore, our theoretical analysis does not exclude the possibility that IDRs themselves may introduce

negative externalities, such as increased risks of abnormal interactions, independent of whether they participate in canonical LLPS.

Although experimental determination of synthetic costs is challenging, our theory generates predictions that can be experimentally tested in future studies. For instance, disruption of the Pigovian taxation of LLPS-related amino acids in human cells, such as expression of high-activity mutants of enzymes in biosynthesis of costly amino acids, could lead to measurable changes in abundance and material property of condensates through lowering the “price” of these amino acids. Changes in the composition of dietary amino acid intake could also perturb the “trading” system, therefore reprogramming central carbon metabolism catalyzed by enzymes predicted to rely more on the trading strategy. It is worth noting that our previous large-scale analysis of amino acid composition of human foods[22] shows that the relative ordering of amino acid abundance in human foods is remarkably robust, supporting the idea that dietary intake of amino acids is constrained to match the composition of conserved central carbon metabolic enzymes.

In summary, our study reveals previously unexplored complexity in the amino-acid economics of multicellular life. In these organisms, the emergence of long-range transport, dietary buffering, and phase-separated organization gives rise to a richer economic landscape in which production, trading, and regulation coexist. Hence, the emergence of AArchetypes highlights a transition from a cost-driven market to a governed economy as a fundamental hallmark of complex life.

Data availability

All datasets used and generated in this study are available at the GitHub page: https://github.com/Alec-Yeung/AArchetypes.

Code availability

All code and scripts generated in this study are available at the GitHub page: https://github.com/Alec-Yeung/AArchetypes.

Author contributions

K.Y. and Z.D. designed the study and wrote the manuscript. K.Y. performed the AArchetype-related analyses. K.S. retrieved and processed the human and yeast proteomics datasets. H.Z. and Y.L. analyzed amino acid sequences of human metabolic enzymes. Z.D. developed the mathematical model for LLPS externality internalization. All authors have read and approved the final manuscript.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (12371489 to ZD and 12205012 to YL), National Key Research and Development Program of China (2021YFA0911300 to ZD), Shenzhen Medical Academy of Research and Translation (D2401022 to ZD), and Basic and Applied Basic Research Foundation of Guangdong Province (2025A1515012923 to YL). The authors thank Prof. Cong Yu, Prof. Kaige Yan, and Prof. Xiandeng Wu for helpful discussions.

Figure legends

Fig. 1. Emergence of AArchetypes in human metabolic enzymes

a: Workflow for calculating amino acid enrichment scores of metabolic pathways.

b: Amino acid enrichment scores of human metabolic pathways and AArchetypes identified by clustering analysis.

c: KEGG metabolic categories enriched in AArchetypes.

d: Receiver operating characteristic (ROC) curves of random forest classifier to assess the separability of the four human AArchetypes using protein-level amino acid composition.

e: Left: workflow for computing pairwise amino acid composition distance and structural distance matrices between proteins; Right: density scatter plot comparing amino acid composition distance to structural distance. Each dot represents a protein; colors indicate point density.

Fig. 2. AArchetypes only emerge in multicellular organisms

a: Phylogenetic tree of 10 unicellular and multicellular organisms.

b: Classical multidimensional scaling (MDS) of pathway-level amino acid composition across 10 organisms. Each dot represents one metabolic pathway from one organism.

c: Three independent metrics assessing the emergence of AArchetype in the ten organisms.

d: Amino acid enrichment scores and silhouette coefficients in metabolic pathways of E. coli (top) and M. musculus (bottom).

Fig. 3. Youngest AArchetype minimizes synthetic cost

a: Schematic overview of the algorithm calculating protein and energy cost for amino acids.

b: Protein cost and energy cost for non-essential amino acids in humans. A two-sided Pearson's correlation test was used to compute the p-value.

c: The consumption or release profile of non-essential amino acids in the NCI-60 human cell line panel. Among these, serine has the highest protein cost, arginine has the highest energy cost, and alanine has the lowest protein cost.

d: Scatter plots comparing proteomic abundance and protein cost across amino acids in yeast (S. cerevisiae, top) and human (H. sapiens, bottom). A two-sided Spearman's rank correlation test was performed to compute the p-values.

e: Violin plots comparing protein cost (left) and energy cost (right) across the four human AArchetypes. P-values were computed using the Kruskal-Wallis test. Bar plots inside the panels show median values for the AArchetypes.
f: Distribution of protein age among the four human AArchetypes. Statistical significance was evaluated using a Chi-square test of independence.
g: Distribution of evolutionary rate for proteins in the four human AArchetypes. Statistical significance was evaluated using the Kruskal-Wallis test.

Fig. 4. Central carbon metabolism AArchetype matches dietary amino acid composition

a: t-distributed neighbor embedding (t-SNE) of amino acid composition profiles of human metabolic pathways (circles for AArchetype A-D) and ten human dietary patterns (triangles).
b: Bubble plot showing the median Euclidean distance between amino acid compositions of each dietary pattern and each AArchetype.

Fig. 5. Costly AArchetypes are associated with structural disorder and phase separation

a: Counts of membrane proteins and other proteins in human AArchetypes. Statistical significance was evaluated using hypergeometric test. Significance: ns: $p \geq 0.05$; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; ****: $p<0.0001$.
b: Violin plot comparing protein abundance across the four human AArchetypes. Statistical significance was evaluated using Wilcoxon's rank-sum test. Significance: ns: $p \geq 0.05$; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; ****: $p<0.0001$.
c: Distribution of Gini coefficients quantifying tissue specificity of protein abundance in AArchetypes. Statistical significance was evaluated using Wilcoxon's rank-sum test. Significance: ns: $p \geq 0.05$; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; ****: $p<0.0001$.

d: Schematic overview of the calculation of protein cost sensitivity scores for human metabolic pathways (top) and the resulting scores (bottom). The numbers above each box plot represent the median value.

e: Distribution of IDR proportion for enzymes in the human AArchetypes. Statistical significance was evaluated using Wilcoxon's rank-sum test. Significance: ns: p≥0.05; *: p<0.05; **: p<0.01; ***: p<0.001; ****: p<0.0001.

f: Counts of LLPS and Non-LLPS proteins in AArchetypes. Statistical significance was evaluated using hypergeometric test.

g: Distribution of normalized enrichment scores (NES) from GSEA for gene sets enriched in costly human proteins. Gene sets include KEGG pathways, and curated sets of IDR- and LLPS-associated proteins.

h: Distribution of protein cost (left) and energy cost (right) of LLPS/non-LLPS protein across the whole human proteome. Statistical significance was evaluated using Wilcoxon's rank-sum test. Significance: ns: p≥0.05; *: p<0.05; **: p<0.01; ***: p<0.001; ****: p<0.0001.

Fig. 6. A mathematical model for LLPS externality internalization by increasing synthetic cost

a: Schematic overview of the mathematical model.

b: Algorithm for estimation of the species-specific LLPS cost ratio w and bar plot showing the values of w in human (*H. sapiens*) and yeast (*S. cerevisiae*).

c: Maximal payoff function $P_{\max}(a)$ under varying a for human (*H. sapiens*).

d: Maximal payoff function $P_{\max}(a)$ under varying a for yeast (*S. cerevisiae*).

e: Optimal solution of the fraction of LLPS-related amino acids a under varying LLPS payoff gain d under the human ($w = 5.77$) and yeast ($w = 0.26$) pricing system.

f: Distribution of the fraction of LLPS-related amino acids a in human proteins.

g: Same as in (f) but for yeast proteins.

h: Distribution of the value of LLPS payoff gain d inferred from the fraction of LLPS-related amino acids a for human proteins.

i: Scatter plot comparing the actual fraction of LLPS-related amino acids a and the simulated optimal a under the yeast-like amino acid pricing regime.

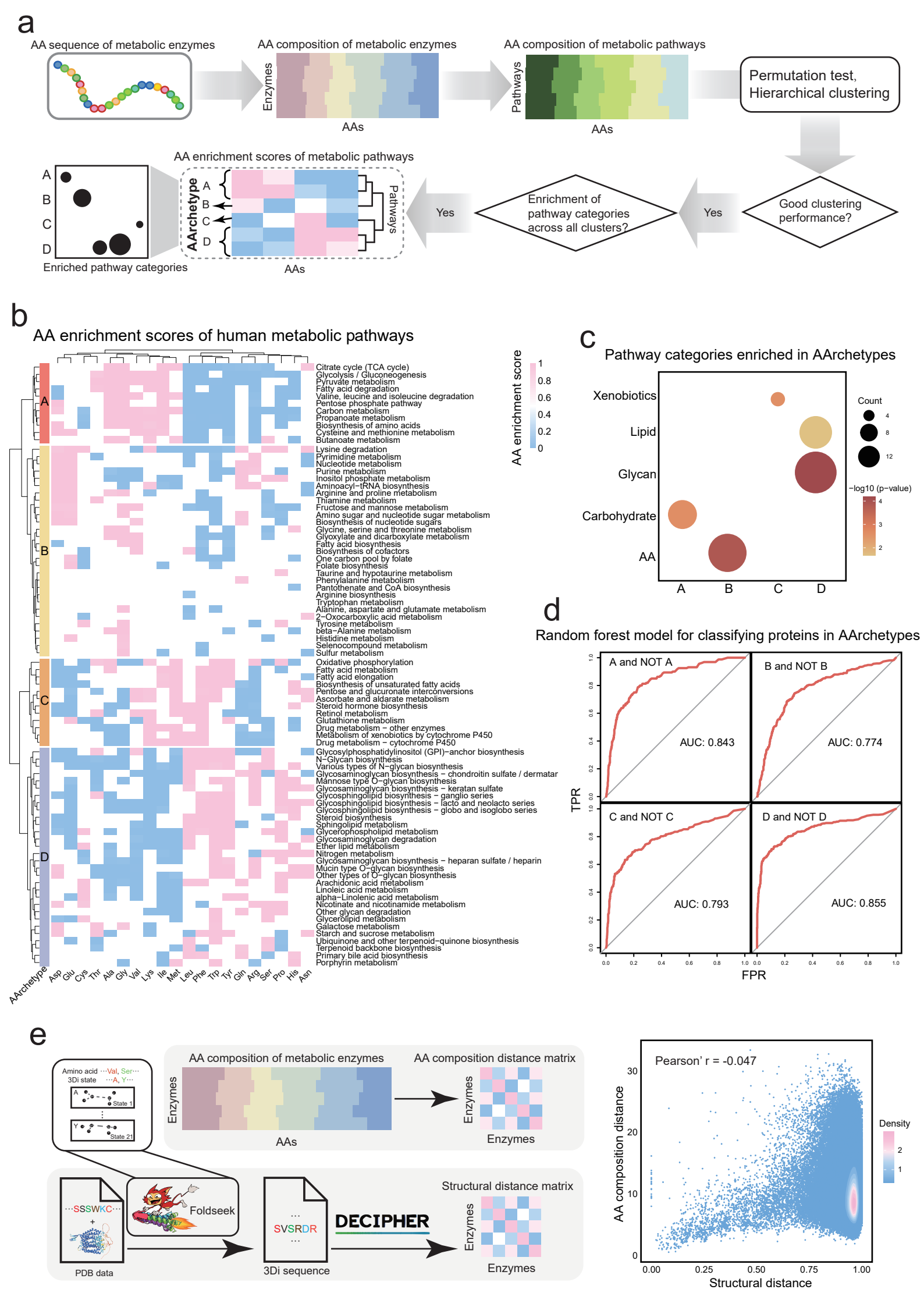
References

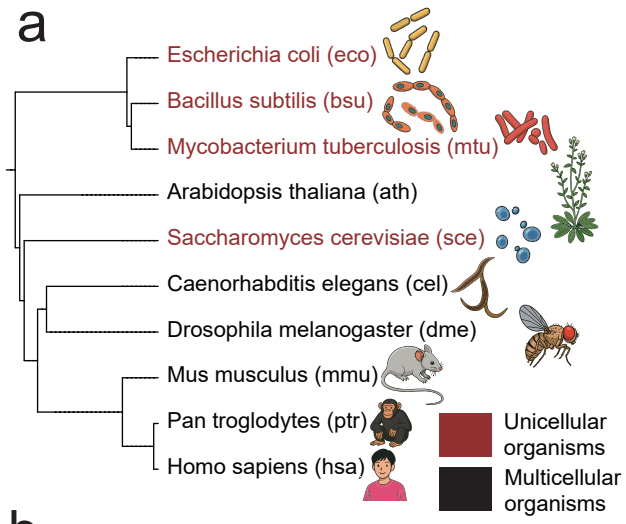
1. Krick, T. et al. Amino Acid metabolism conflicts with protein diversity. *Mol Biol Evol* **31**, 2905-2912 (2014).
2. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* **99**, 3695-3700 (2002).
3. Swire, J. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol* **64**, 558-571 (2007).
4. Chen, Y. & Nielsen, J. Yeast has evolved to minimize protein resource cost for synthesizing amino acids. *Proc Natl Acad Sci U S A* **119** (2022).
5. Zhang, H. et al. Biosynthetic energy cost for amino acids decreases in cancer evolution. *Nat Commun* **9**, 4124 (2018).
6. Zeldovich, K.B., Berezovsky, I.N. & Shakhnovich, E.I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* **3**, e5 (2007).
7. Flamholz, A.I., Goyal, A., Fischer, W.W., Newman, D.K. & Phillips, R. The proteome is a terminal electron acceptor. *Proc Natl Acad Sci U S A* **122**, e2404048121 (2025).
8. Mende, D.R. et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol* **2**, 1367-1373 (2017).
9. Piper, M.D.W. et al. Matching Dietary Amino Acid Balance to the In Silico-Translated Exome Optimizes Growth and Reproduction without Cost to Lifespan. *Cell Metab* **25**, 1206 (2017).
10. Grandison, R.C., Piper, M.D. & Partridge, L. Amino-acid imbalance explains extension of lifespan by dietary restriction in Drosophila. *Nature* **462**, 1061-1064 (2009).
11. Guedes, R.L. et al. Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution. *BMC Genomics* **12 Suppl 4**, S2 (2011).
12. Raiford, D.W. et al. Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? *J Mol Evol* **67**, 621-630 (2008).
13. Fleming, J. et al. AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *J Mol Biol* **437**, 168967 (2025).
14. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243-246 (2024).

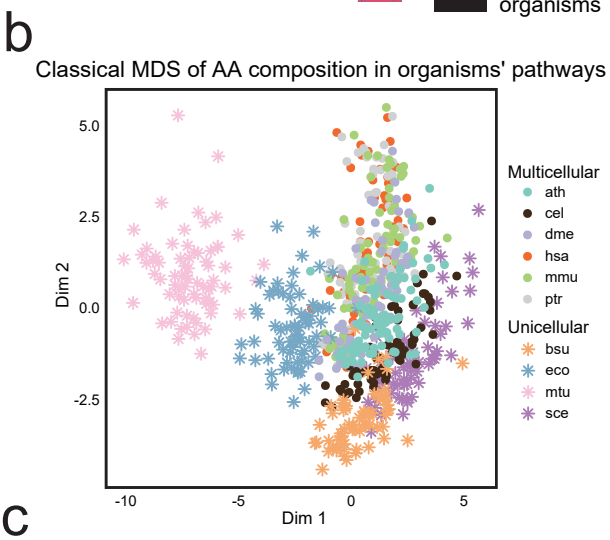
15. Foreman, J. et al. DECIPHER: Improving Genetic Diagnosis Through Dynamic Integration of Genomic and Clinical Data. *Annu Rev Genomics Hum Genet* **24**, 151-176 (2023).
16. Lautenbacher, L. et al. ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Res* **50**, D1541-D1552 (2022).
17. Nusinow, D.P. et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387-402 e316 (2020).
18. Ho, B., Baryshnikova, A. & Brown, G.W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst* **6**, 192-205 e193 (2018).
19. Hosios, A.M. et al. Amino Acids Rather than Glucose Account for the Majority of Cell Mass in Proliferating Mammalian Cells. *Dev Cell* **36**, 540-549 (2016).
20. Shao, Y. et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res* **29**, 682-696 (2019).
21. Zhang, J. & Yang, J.R. Determinants of the rate of protein sequence evolution. *Nat Rev Genet* **16**, 409-420 (2015).
22. Dai, Z., Zheng, W. & Locasale, J.W. Amino acid variability, tradeoffs and optimality in human diet. *Nat Commun* **13**, 6683 (2022).
23. Chen, Y. et al. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. *Nat Protoc* **19**, 629-667 (2024).
24. Lotthammer, J.M., Ginell, G.M., Griffith, D., Emenecker, R.J. & Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods* **21**, 465-476 (2024).
25. Ning, W. et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* **48**, D288-D295 (2020).
26. Liebermeister, W. et al. Visual account of protein investment in cellular functions. *Proc Natl Acad Sci U S A* **111**, 8488-8493 (2014).
27. Arkin, I.T. & Brunger, A.T. Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta* **1429**, 113-128 (1998).
28. White, S.H. & Wimley, W.C. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* **28**, 319-365 (1999).
29. Radivojac, P. et al. Intrinsic disorder and functional proteomics. *Biophys J* **92**, 1439-1456 (2007).
30. Alberti, S. & Dormann, D. Liquid-Liquid Phase Separation in Disease. *Annu Rev Genet* **53**, 171-194 (2019).

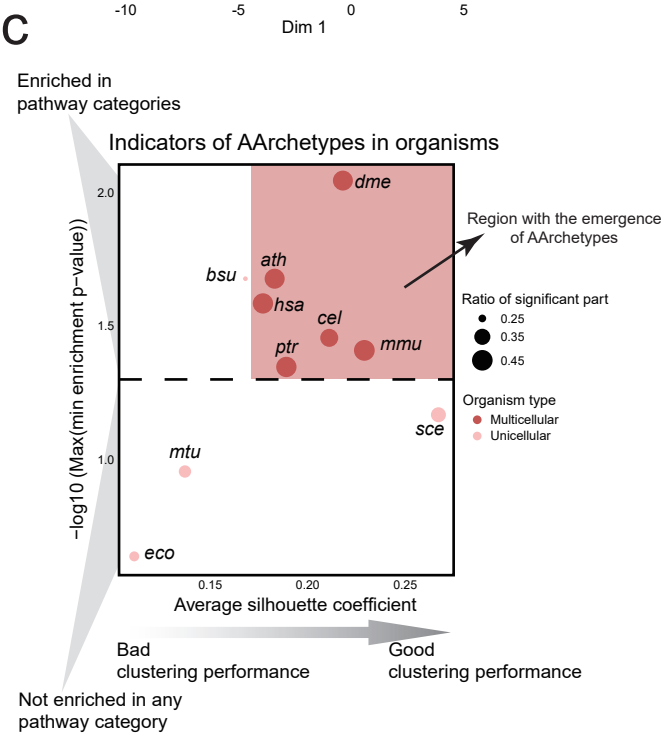
31. An, S., Kumar, R., Sheets, E.D. & Benkovic, S.J. Reversible compartmentalization of de novo purine biosynthetic complexes in living cells. *Science* **320**, 103-106 (2008).
32. Castellana, M. et al. Enzyme clustering accelerates processing of intermediates through metabolic channeling. *Nat Biotechnol* **32**, 1011-1018 (2014).
33. Peeples, W. & Rosen, M.K. Mechanistic dissection of increased enzymatic rate in a phase-separated compartment. *Nat Chem Biol* **17**, 693-702 (2021).
34. O'Flynn, B.G. & Mittag, T. The role of liquid-liquid phase separation in regulating enzyme activity. *Curr Opin Cell Biol* **69**, 70-79 (2021).
35. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. & Sharp, P.A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23 (2017).
36. Shin, Y. & Brangwynne, C.P. Liquid phase condensation in cell physiology and disease. *Science* **357** (2017).
37. Patel, A. et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**, 1066-1077 (2015).
38. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D.S. How protein stability and new functions trade off. *PLoS Comput Biol* **4**, e1000002 (2008).
39. Heaton, L.L.M., Jones, N.S. & Fricker, M.D. A mechanistic explanation of the transition to simple multicellularity in fungi. *Nat Commun* **11**, 2594 (2020).
40. Pigou, A.C. The economics of welfare. (Macmillan and co., ltd., London,; 1920).
41. Vavouri, T., Semple, J.I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198-208 (2009).
42. Yang, L. et al. Phase separation as a possible mechanism for dosage sensitivity. *Genome Biol* **25**, 17 (2024).
43. Bolognesi, B. et al. A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* **16**, 222-231 (2016).
44. van der Lee, R. et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep* **8**, 1832-1844 (2014).
45. Jeon, S. et al. Emerging regulatory mechanisms and functions of biomolecular condensates: implications for therapeutic targets. *Signal Transduct Target Ther* **10**, 4 (2025).
46. Sridharan, S. et al. Systematic discovery of biomolecular condensate-specific protein phosphorylation. *Nat Chem Biol* **18**, 1104-1114 (2022).
47. Ranganathan, S., Dasmeh, P., Furniss, S. & Shakhnovich, E. Phosphorylation sites are evolutionary checkpoints against liquid-solid

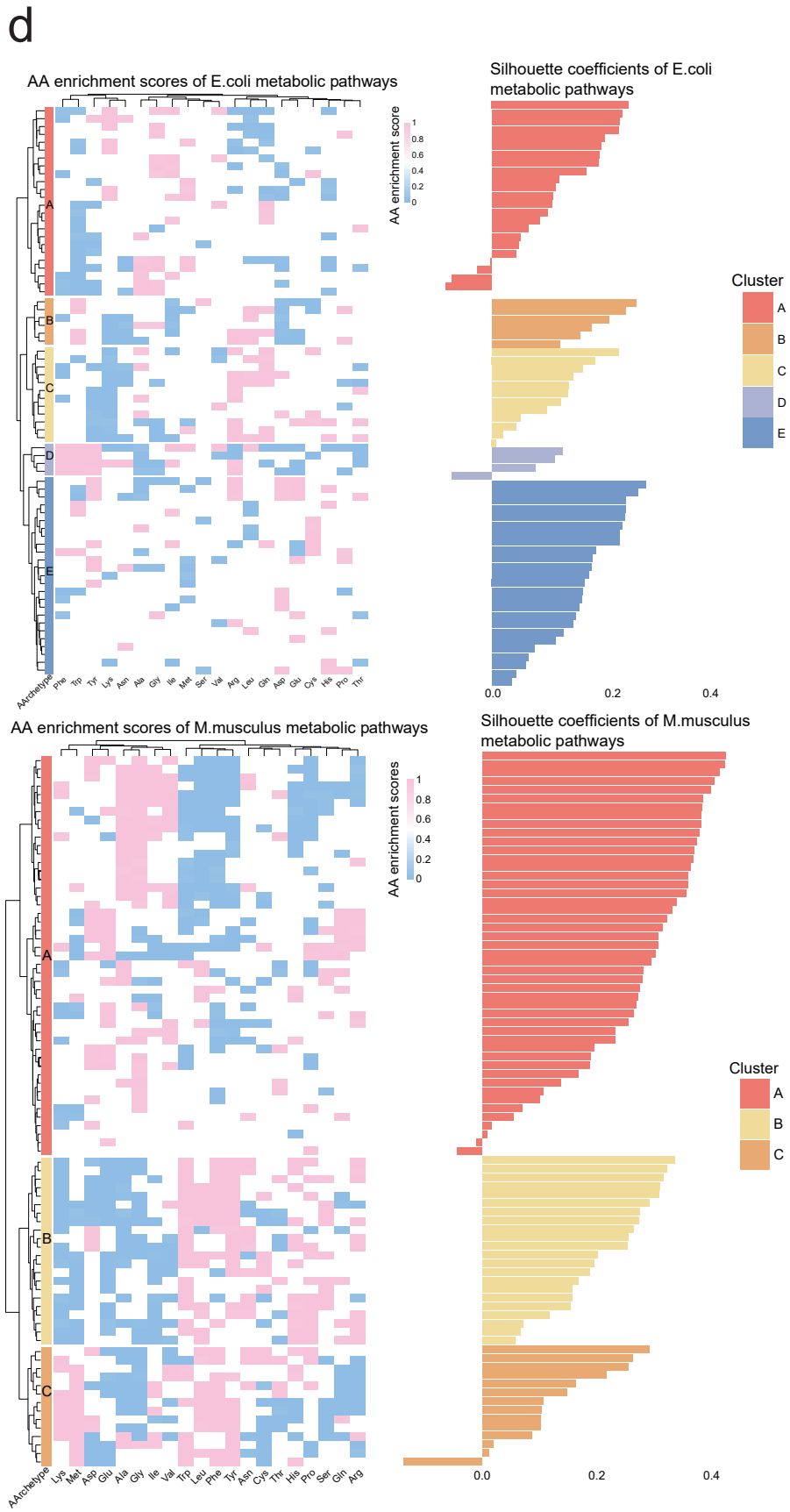
transition in protein condensates. *Proc Natl Acad Sci U S A* **120**, e2215828120 (2023).
48. Wilfling, F. et al. A Selective Autophagy Pathway for Phase-Separated Endocytic Protein Deposits. *Mol Cell* **80**, 764-778 e767 (2020).
49. Shao, Y. et al. A chaperone-like function of FUS ensures TAZ condensate dynamics and transcriptional activation. *Nat Cell Biol* **26**, 86-99 (2024).
50. Stoffel, F. et al. Enhancement of enzymatic activity by biomolecular condensates through pH buffering. *Nat Commun* **16**, 6368 (2025).
51. Wang, J. et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688-699 e616 (2018).
52. Wiedner, H.J. & Giudice, J. It's not just a phase: function and characteristics of RNA-binding proteins in phase separation. *Nat Struct Mol Biol* **28**, 465-473 (2021).
53. Molliex, A. et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123-133 (2015).
54. Dunker, A.K. et al. Intrinsically disordered protein. *J Mol Graph Model* **19**, 26-59 (2001).
55. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**, 2-8 (2008).
56. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. & Dunker, A.K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**, 573-584 (2002).

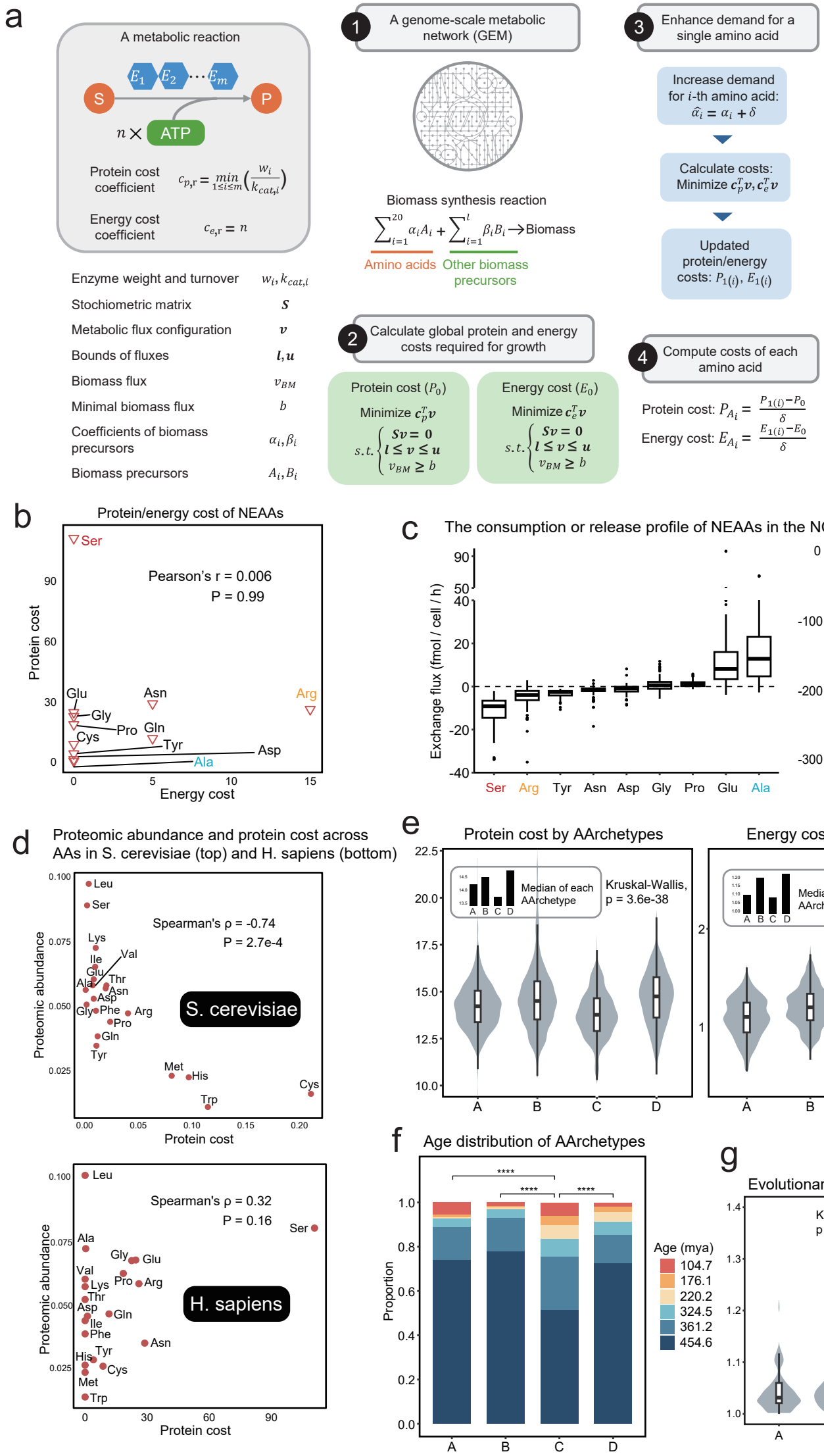

a
AA sequence of metabolic enzymes
AA composition of metabolic enzymes
Enzymes
AAs
AA composition of metabolic pathways
Pathways
AAs
Permutation test, Hierarchical clustering
Good clustering performance?
Yes
Enrichment of pathway categories across all clusters?
Yes
AA enrichment scores of metabolic pathways
AArchetype
A
B
C
D
Pathways
AAs
Enriched pathway categories
b
AA enrichment scores of human metabolic pathways
AA enrichment score
1
0.8
0.6
0.4
0.2
0
Citrate cycle (TCA cycle)
Glycolysis / Gluconeogenesis
Pyruvate metabolism
Fatty acid degradation
Valine, leucine and isoleucine degradation
Pentose phosphate pathway
Carbon metabolism
Propanoate metabolism
Biosynthesis of amino acids
Cysteine and methionine metabolism
Butanoate metabolism
Lysine degradation
Pyrimidine metabolism
Nucleotide metabolism
Purine metabolism
Inositol phosphate metabolism
Aminoacyl−tRNA biosynthesis
Arginine and proline metabolism
Thiamine metabolism
Fructose and mannose metabolism
Amino sugar and nucleotide sugar metabolism
Biosynthesis of nucleotide sugars
Glycine, serine and threonine metabolism
Glyoxylate and dicarboxylate metabolism
Fatty acid biosynthesis
Biosynthesis of cofactors
One carbon pool by folate
Folate biosynthesis
Taurine and hypotaurine metabolism
Phenylalanine metabolism
Pantothenate and CoA biosynthesis
Arginine biosynthesis
Tryptophan metabolism
Alanine, aspartate and glutamate metabolism
2−Oxocarboxylic acid metabolism
Tyrosine metabolism
beta−Alanine metabolism
Histidine metabolism
Selenocompound metabolism
Sulfur metabolism
Oxidative phosphorylation
Fatty acid metabolism
Fatty acid elongation
Biosynthesis of unsaturated fatty acids
Pentose and glucuronate interconversions
Ascorbate and aldarate metabolism
Steroid hormone biosynthesis
Retinol metabolism
Glutathione metabolism
Drug metabolism − other enzymes
Metabolism of xenobiotics by cytochrome P450
Drug metabolism − cytochrome P450
Glycosylphosphatidylinositol (GPI)−anchor biosynthesis
N−Glycan biosynthesis
Various types of N−glycan biosynthesis
Glycosaminoglycan biosynthesis − chondroitin sulfate / dermatar
Mannose type O−glycan biosynthesis
Glycosaminoglycan biosynthesis − keratan sulfate
Glycosphingolipid biosynthesis − ganglio series
Glycosphingolipid biosynthesis − lacto and neolacto series
Glycosphingolipid biosynthesis − globo and isoglobo series
Steroid biosynthesis
Sphingolipid metabolism
Glycerophospholipid metabolism
Glycosaminoglycan degradation
Ether lipid metabolism
Nitrogen metabolism
Glycosaminoglycan biosynthesis − heparan sulfate / heparin
Mucin type O−glycan biosynthesis
Other types of O−glycan biosynthesis
Arachidonic acid metabolism
Linoleic acid metabolism
alpha−Linolenic acid metabolism
Nicotinate and nicotinamide metabolism
Other glycan degradation
Glycerolipid metabolism
Galactose metabolism
Starch and sucrose metabolism
Ubiquinone and other terpenoid−quinone biosynthesis
Terpenoid backbone biosynthesis
Primary bile acid biosynthesis
Porphyrin metabolism
AArchetype
Asp Glu Cys Thr Ala Gly Val Lys Ile Met Leu Phe Trp Tyr Gln Arg Ser Pro His Asn
c
Pathway categories enriched in AArchetypes
Xenobiotics
Lipid
Glycan
Carbohydrate
AA
A B C D
Count
4
8
12
−log10 (p−value)
4
3
2
d
Random forest model for classifying proteins in AArchetypes
A and NOT A
AUC: 0.843
B and NOT B
AUC: 0.774
C and NOT C
AUC: 0.793
D and NOT D
AUC: 0.855
TPR
FPR
0.0 0.2 0.4 0.6 0.8 1.0
e
Amino acid ···Val, Ser···
3Di state ···A, Y···
State 1
State 21
AA composition of metabolic enzymes
Enzymes
AAs
AA composition distance matrix
Enzymes
Enzymes
···SSSWKC···
Foldseek
PDB data
SVSRDR
3Di sequence
DECIPHER
Structural distance matrix
Enzymes
Enzymes
Pearson' r = -0.047
AA composition distance
Structural distance
0.00 0.25 0.50 0.75 1.00
0 10 20 30
Density
2
1

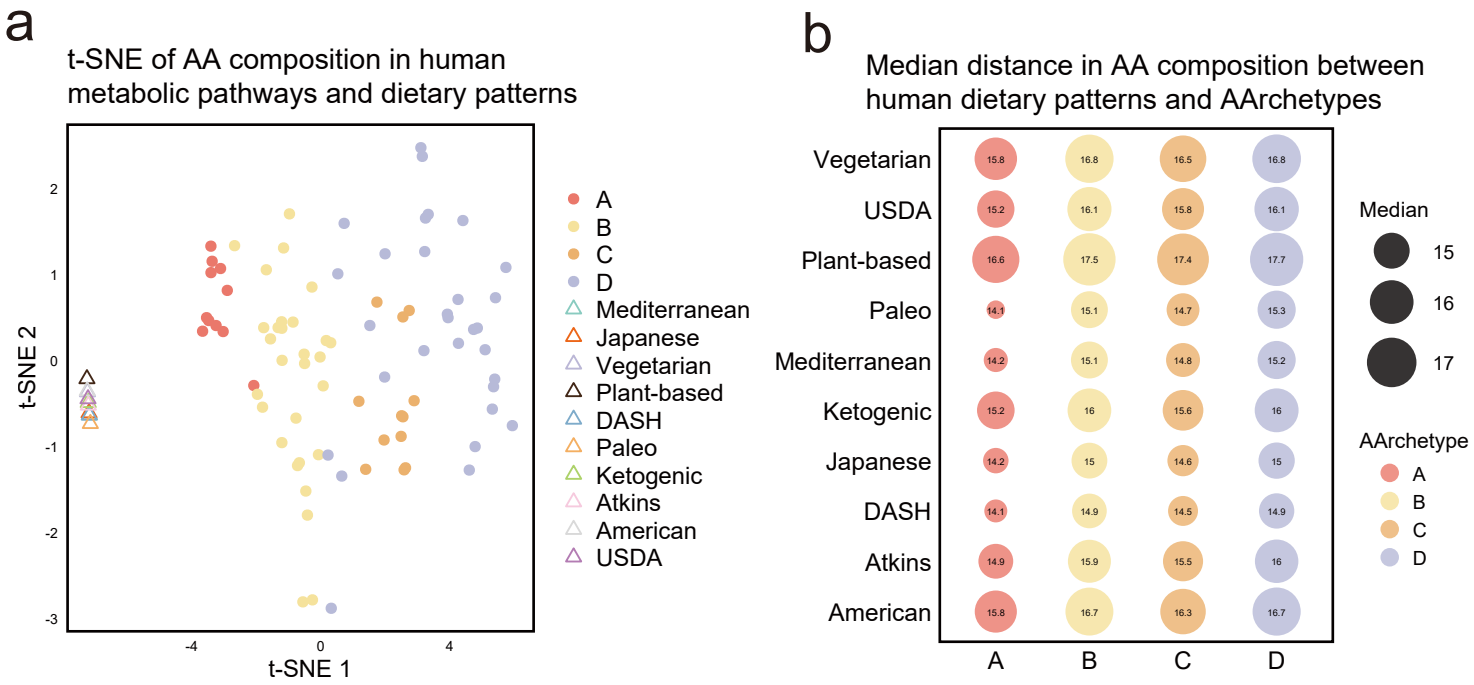

a
Escherichia coli (eco)
Bacillus subtilis (bsu)
Mycobacterium tuberculosis (mtu)
Arabidopsis thaliana (ath)
Saccharomyces cerevisiae (sce)
Caenorhabditis elegans (cel)
Drosophila melanogaster (dme)
Mus musculus (mmu)
Pan troglodytes (ptr)
Homo sapiens (hsa)
Unicellular organisms
Multicellular organisms

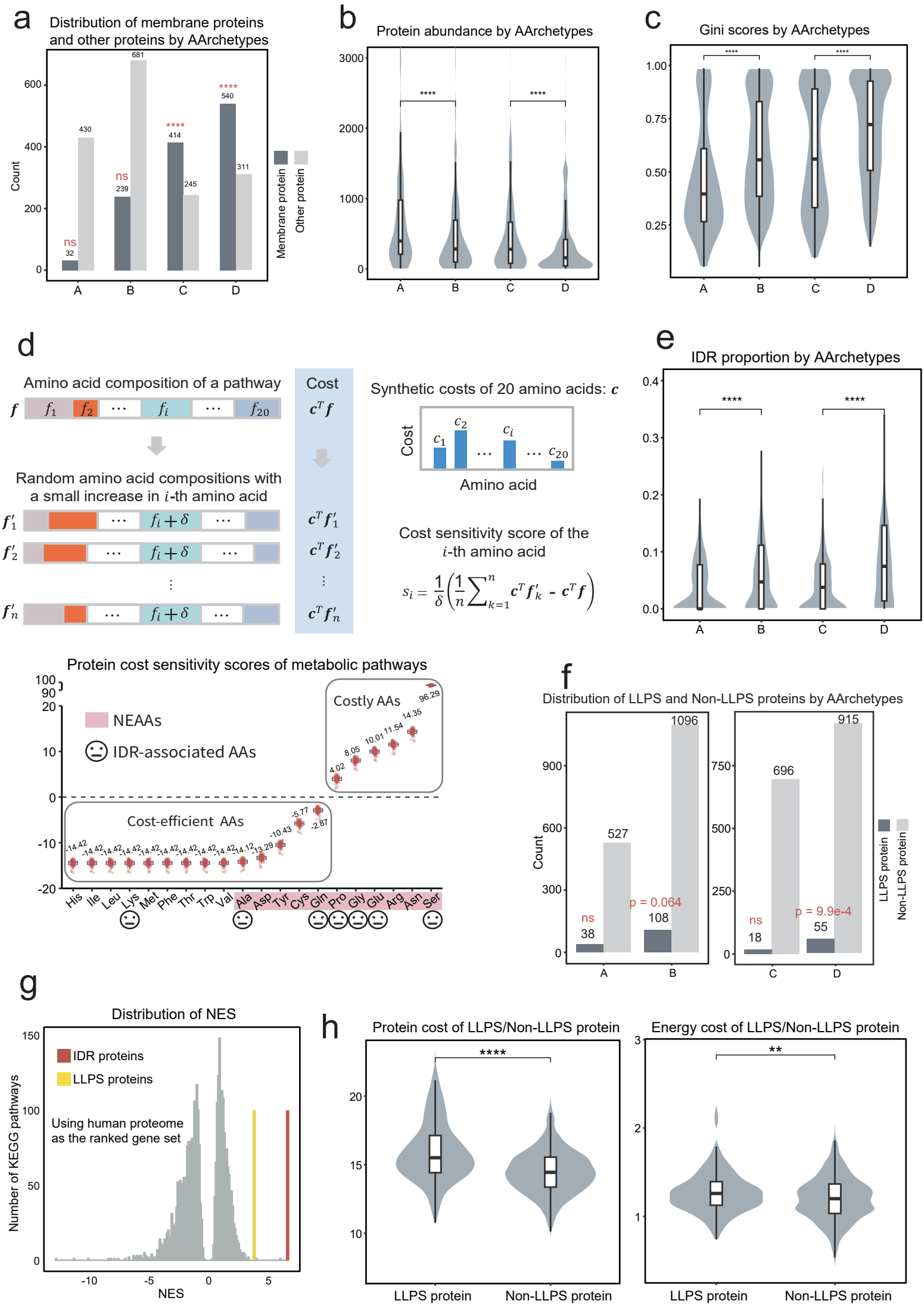

b
Classical MDS of AA composition in organisms' pathways
Dim 2
Dim 1
5.0
2.5
0.0
-2.5
-10
-5
0
5
Multicellular
ath
cel
dme
hsa
mmu
ptr
Unicellular
bsu
eco
mtu
sce


c
Enriched in pathway categories
Indicators of AArchetypes in organisms
−log10 (Max(min enrichment p−value))
Average silhouette coefficient
2.0
1.5
1.0
0.15
0.20
0.25
dme
ath
bsu
hsa
cel
ptr
mmu
sce
mtu
eco
Region with the emergence of AArchetypes
Ratio of significant part
0.25
0.35
0.45
Organism type
Multicellular
Unicellular
Bad clustering performance
Good clustering performance
Not enriched in any pathway category


d
AA enrichment scores of E.coli metabolic pathways
AA enrichment score
Silhouette coefficients of E.coli metabolic pathways
Cluster
A
B
C
D
E
AArchetype
Phe Trp Tyr Lys Asn Ala Gly Ile Met Ser Val Arg Leu Gln Asp Glu Cys His Pro Thr
0.0
0.2
0.4
AA enrichment scores of M.musculus metabolic pathways
AA enrichment scores
Silhouette coefficients of M.musculus metabolic pathways
Cluster
A
B
C
AArchetype
Lys Met Asp Glu Ala Gly Ile Val Trp Leu Phe Tyr Asn Cys Thr His Pro Ser Gln Arg
0.0
0.2
0.4

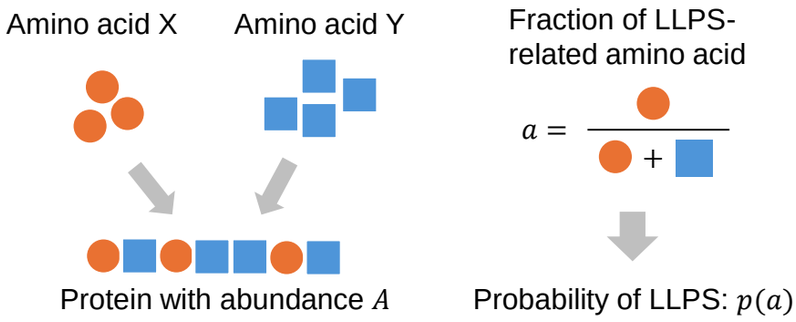

a
A metabolic reaction
S
P
ATP
Protein cost coefficient
Energy cost coefficient
Enzyme weight and turnover
Stochiometric matrix
Metabolic flux configuration
Bounds of fluxes
Biomass flux
Minimal biomass flux
Coefficients of biomass precursors
Biomass precursors
1 A genome-scale metabolic network (GEM)
Biomass synthesis reaction
Amino acids
Other biomass precursors
2 Calculate global protein and energy costs required for growth
Protein cost (P_0)
Energy cost (E_0)
3 Enhance demand for a single amino acid
Increase demand for i-th amino acid:
Calculate costs:
Updated protein/energy costs:
4 Compute costs of each amino acid
Protein cost:
Energy cost:
b
Protein/energy cost of NEAAs
Pearson's r = 0.006
P = 0.99
Protein cost
Energy cost
c
The consumption or release profile of NEAAs in the NCI-60 cell line
Exchange flux (fmol / cell / h)
Ser
Arg
Tyr
Asn
Asp
Gly
Pro
Glu
Ala
Gln
d
Proteomic abundance and protein cost across AAs in S. cerevisiae (top) and H. sapiens (bottom)
Spearman's ρ = -0.74
P = 2.7e-4
S. cerevisiae
Spearman's ρ = 0.32
P = 0.16
H. sapiens
Proteomic abundance
Protein cost
e
Protein cost by AArchetypes
Median of each AArchetype
Kruskal-Wallis, p = 3.6e-38
Energy cost by AArchetypes
Kruskal-Wallis, p = 6.5e-39
f
Age distribution of AArchetypes
Proportion
Age (mya)
104.7
176.1
220.2
324.5
361.2
454.6
g
Evolutionary rate by AArchetypes
Kruskal-Wallis, p = 5.9e-19


a
t-SNE of AA composition in human metabolic pathways and dietary patterns
t-SNE 2
t-SNE 1
2
1
0
-1
-2
-3
-4
0
4
A
B
C
D
Mediterranean
Japanese
Vegetarian
Plant-based
DASH
Paleo
Ketogenic
Atkins
American
USDA
b
Median distance in AA composition between human dietary patterns and AArchetypes
Vegetarian 15.8 16.8 16.5 16.8
USDA 15.2 16.1 15.8 16.1
Plant-based 16.6 17.5 17.4 17.7
Paleo 14.1 15.1 14.7 15.3
Mediterranean 14.2 15.1 14.8 15.2
Ketogenic 15.2 16 15.6 16
Japanese 14.2 15 14.6 15
DASH 14.1 14.9 14.5 14.9
Atkins 14.9 15.9 15.5 16
American 15.8 16.7 16.3 16.7
A
B
C
D
Median
15
16
17
AArchetype
A
B
C
D


a
Distribution of membrane proteins and other proteins by AArchetypes
Count
Membrane protein
Other protein
ns
****
b
Protein abundance by AArchetypes
c
Gini scores by AArchetypes
d
Amino acid composition of a pathway
Cost
Random amino acid compositions with a small increase in i-th amino acid
Synthetic costs of 20 amino acids: c
Amino acid
Cost sensitivity score of the i-th amino acid
$s_i = \frac{1}{\delta}\left(\frac{1}{n}\sum_{k=1}^{n} c^T f'_k - c^T f\right)$
e
IDR proportion by AArchetypes
Protein cost sensitivity scores of metabolic pathways
NEAAs
IDR-associated AAs
Costly AAs
Cost-efficient AAs
His Ile Leu Lys Met Phe Thr Trp Val Ala Asp Tyr Cys Gln Pro Gly Glu Arg Asn Ser
f
Distribution of LLPS and Non-LLPS proteins by AArchetypes
p = 0.064
p = 9.9e-4
LLPS protein
Non-LLPS protein
g
Distribution of NES
IDR proteins
LLPS proteins
Using human proteome as the ranked gene set
Number of KEGG pathways
NES
h
Protein cost of LLPS/Non-LLPS protein
Energy cost of LLPS/Non-LLPS protein
LLPS protein
Non-LLPS protein

a

Group	LLPS-related	Average synthetic cost	Amino acids
X (●)	Yes	w	K, A, Q, P, G, E, S
Y (■)	No	1	R, N, D, C, H, I, L, M, F, T, W, Y, V



Payoff function: $P(a, A) = P_{\text{func}}(a, A) + \beta P_{\text{info}}(a)$

Functional benefit — Information capacity

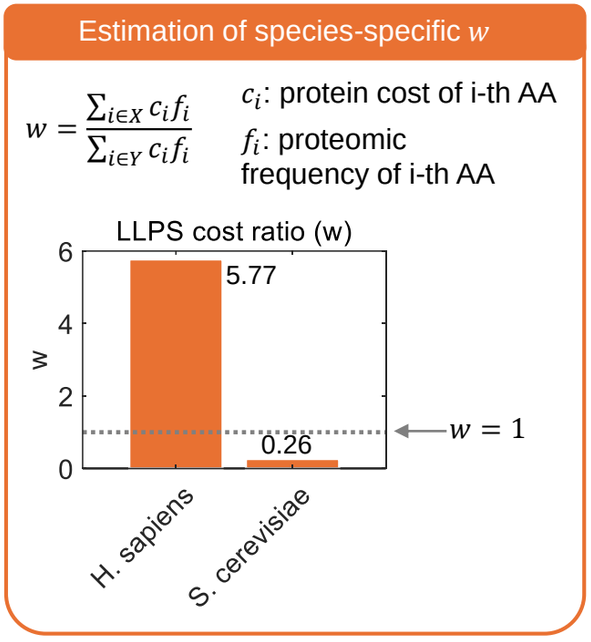
$P_{\text{func}}(a, A) = A(1 + dp(a))$

$P_{\text{info}}(a) = -a \log_2 \frac{a}{7} - (1-a) \log_2 \frac{1-a}{13}$

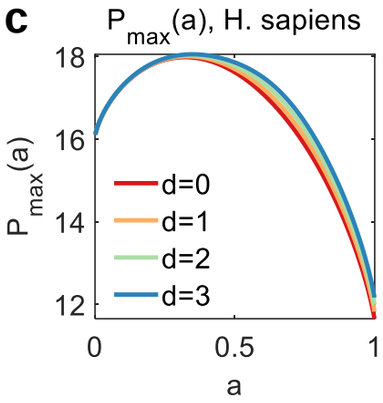
Maximize $P(a, A)$

$s.t.\ (wa + (1-a))A \leq 1$

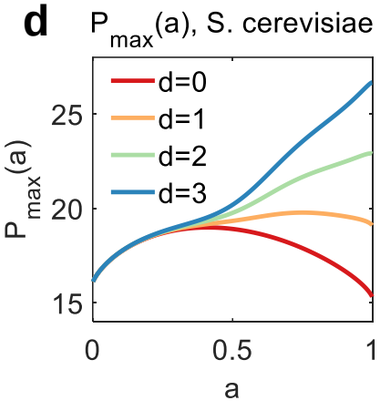
b



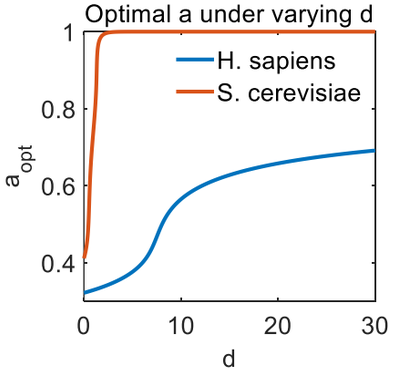
c



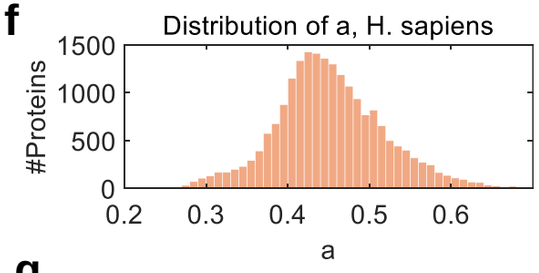
d



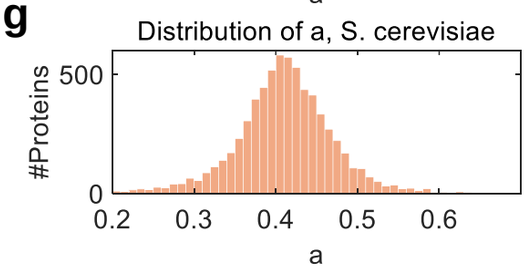
e



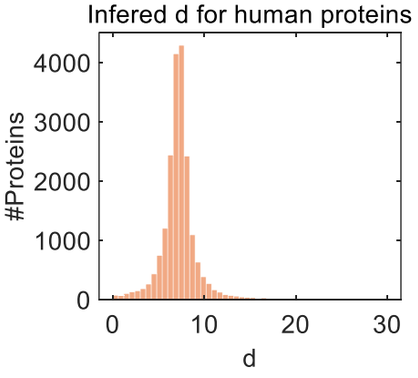
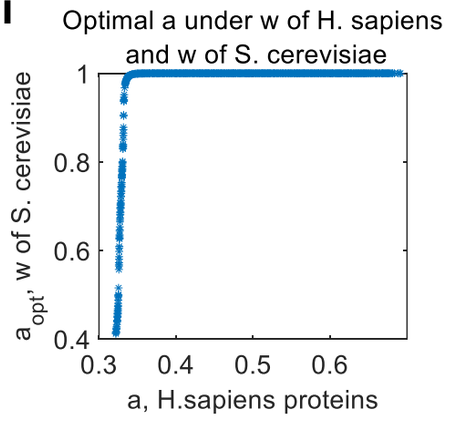
f



g



h



i