Revealing global stoichiometry conservation architecture in cells from Raman spectral patterns

Ken-ichiro F. Kamei,[1*] Koseki J. Kobayashi-Kirschvink,[2]
Takashi Nozoe,[1,3,4] Hidenori Nakaoka,[5] Miki Umetani,[6]
Yuichi Wakamoto[1,3,4*]

[1]Department of Basic Science, Graduate School of Arts and Sciences,
The University of Tokyo, Tokyo 153-8902, Japan
[2]Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
[3]Research Center for Complex Systems Biology, The University of Tokyo,
Tokyo 153-8902, Japan
[4]Universal Biology Institute, The University of Tokyo, Tokyo 153-8902, Japan
[5]Department of Optical Imaging, Advanced Research Promotion Center
Tokushima University, Tokushima 770-8503, Japan
[6]Department of Biology, New York University, New York, NY 10003, USA

*To whom correspondence should be addressed; E-mail:
kenichiro_kamei@cell.c.u-tokyo.ac.jp (K.F.K.), cwaka@mail.ecc.u-tokyo.ac.jp (Y.W.).

Cells can adapt to various environments by changing their biomolecular profiles while maintaining physiological homeostasis. What organizational principles in cells enable the simultaneous realization of adaptability and homeostasis? To address this question, we measure Raman scattering light from *Escherichia coli* cells under diverse conditions, whose spectral patterns convey their comprehensive molecular composition. We reveal that dimension-reduced Raman spectra can predict condition-dependent proteome profiles. Quantitative analysis of the Raman-proteome correspondence characterizes a low-dimensional hierarchical stoichiometry-conserving proteome structure. The network centrality of each gene in the stoichiometry conservation relations correlates with its essentiality and evolutionary conservation, and these correlations are preserved from bacteria to human cells. Furthermore, stoichiometry-conserving core components obey growth law and ensure homeostasis across conditions, whereas peripheral stoichiometry-conserving components enable adaptation to specific conditions. Mathematical analysis reveals that the stoichiometrically constrained architecture is reflected in major changes in Raman spectral patterns. These results uncover coordination of global stoichiometric balance in cells and demonstrate that vibrational spectroscopy can decipher such biological constraints beyond statistical or machine-learning inference of cellular states.

Introduction

Biological cells can change their gene expression and metabolic profiles globally to adapt to thier biological contexts and external conditions, while maintaining the homeostasis of their core physiological states. The simultaneous realization of adaptability and homeostasis is a hallmark of biological systems and is assumed to be a system-level property of gene expression profiles in cells (*1, 2*). However, understanding the underlying organizational principles in comprehensive gene expression profiles remains to be a fundamental problem in biology.

Vibrational spectroscopy such as Raman spectroscopy might help us investigate such principles in gene expression profiles. Raman spectroscopy is a light scattering technique that measures energy shifts of light caused by interaction with sample molecules. Raman spectra are obtainable non-destructively even from biological samples such as individual cells. In principle, cellular Raman spectra are optical signatures conveying comprehensive molecular composition of targeted cells (*3–6*). Furthermore, no prior treatments, such as staining and tagging, are necessary to obtain cellular Raman spectra. However, although some biomolecules have separable and intense Raman signal peaks, Raman spectra of most biomolecules overlap and are masked by signals of other molecules due to the diversity and complexity of molecular compositions of cells. Therefore, it is impractical to comprehensively determine the amounts of biomolecules by spectral decomposition.

Despite the intractability of spectral decomposition, reconstruction of comprehensive molecular profiles may be achievable by analyzing detectable global spectral patterns (Fig. 1A) thanks to effective low dimensionality of changes in molecular profile of targeted cells (*7–17*) (Fig. 1B and Fig. S1). Indeed, it has been demonstrated that condition-dependent global transcriptome profiles of cells can be inferred from cellular Raman spectra based on their statistical correspondence (*18, 19*). Importantly, this Raman-spectroscopic transcriptome inference was pos-

sible from dimension-reduced Raman spectra. Therefore, dominant changes in global Raman spectral patterns may contain vital information about the constraints on the molecular profiles in cells; an inspection of their correspondence might give us insights into architectural principles of omics profiles and biological foundation for global omics inference from spectral patterns (Fig. S1).

In this report, we first reveal that, in addition to transcriptomes, condition-dependent proteome profiles of *Escherichia coli* are predictable from cellular Raman spectra. Next, we scrutinize the correspondence between Raman and proteome data, identifying several stoichiometrically conserved groups (SCGs) whose expression tightly correlates with the major changes in cellular Raman spectra. Finally, we reveal that the stoichiometry conservation centrality of each gene correlates with its essentiality, evolutionary conservation, and condition-specificity of gene expression levels, which turns out general across different omics layers and organisms.

RESULTS

Statistical correspondence between Raman spectra and proteomes To examine the correspondence between Raman spectra and proteomes in *E. coli*, we reproduced 15 environmental conditions for which absolute quantitative proteome data are already available (*20*) and measured Raman spectra of *E. coli* cells under those conditions (Fig. 2A and Fig. 2B). The culture conditions we adopted include (i) exponential growth phase in minimal media with various carbon sources, (ii) exponential growth phase in rich media, (iii) exponential growth phase with various stressors, and (iv) stationary phases (Table S1). We measured Raman spectra of single cells sampled from each condition and focused on the fingerprint region of biological samples, where the signals from various biomolecules concentrate (spectral range of 700–1800 cm^{-1}, Fig. 2B and Fig. S2). The Raman spectra were classified on the basis of the environmental conditions using a simple linear classifier, linear discriminant analysis (LDA) (*3, 4, 21*) (Fig.

2C–2E and S1). This classifier calculates the most discriminatory axes by maximizing the ratio of between-condition variance to within-condition variance and reduces the dimensions of Raman data to $m-1$, where $m = 15$ is the number of conditions (see sections 1.1 and 2.1 in (*22*)).

The result shows that Raman spectral points from different environmental conditions are distinguishable in the $(m-1)$-dimensional LDA space (Fig. 2C–2E). For example, the first and second LDA axes clearly distinguish the conditions "LB" and "stationary3days" (Fig. 2C), and the third axis distinguishes "Glucose42C" and "GlycerolAA" (Fig. 2D). Notably, the first principal axis LDA1 correlated with growth rate significantly (Pearson correlation $r = 0.81 \pm 0.09$, Fig. S2). Visualizing the Raman LDA data by embedding them on a 2D plane using t-distributed Stochastic Neighbor Embedding (t-SNE) (*23*) confirms that the points for each condition form a distinctive cluster (Fig. 2F). These results imply that positions in the Raman LDA space reflect condition-dependent differences in cellular physiological states.

We next asked whether these Raman spectral differences in the LDA space could be linked to the different proteome profiles (Fig. S1). To examine this, we hypothesized linear correspondence between the n-dimensional proteome column vector $\hat{\mathbf{p}}_j$, where $n = 2058$ is the number of protein species in the proteome data, and the low-dimensional ($(m-1)$-dimensional) Raman column vector $\hat{\mathbf{r}}_j$ in condition j,

$$\hat{\mathbf{p}}_j = \mathrm{B} \cdot \begin{bmatrix} 1 \\ \hat{\mathbf{r}}_j \end{bmatrix}. \tag{1}$$

B is an $n \times m$ matrix that connects $\hat{\mathbf{p}}_j$ and $\hat{\mathbf{r}}_j$. We calculated $\hat{\mathbf{r}}_j$ as the average of the low-dimensional LDA Raman data of single cells in condition j since the proteomes were measured for cell populations (Table 1).

We conducted leave-one-out cross-validation (LOOCV) to verify this linear correspondence (Fig. 2G). We excluded one condition (here, j) as a test condition and estimated B as $\mathrm{B}^{\mathrm{est}}_{-j}$ by simple ordinary least squares (OLS) regression using the data of the rest of the conditions. We

thereby estimated the proteome in condition j as $\hat{\mathbf{p}}_j^{\mathrm{est}} = \mathrm{B}_{-j}^{\mathrm{est}} \cdot \begin{bmatrix} 1 \\ \hat{\mathbf{r}}_j \end{bmatrix}$.

The proteome profile estimated using the first four major LDA axes (LDA1–LDA4) agreed well with the actual proteome data under most conditions (Fig. 2H and Fig. S3; see section 1.2 in (*22*) for the estimation with all the LDA axes). Changing the condition to exclude, we estimated the proteomes for all the 15 conditions and calculated the overall estimation error by the Euclidean distance $\sum_j \|\hat{\mathbf{p}}_j^{\mathrm{est}} - \hat{\mathbf{p}}_j\|^2$. The result shows that the overall estimation error is significantly small ($p = 0.00005$ by permutation test (*24–26*)). Adopting other distance measures does not change the conclusion (Table S2 and Table S3). These results therefore validate the assumption of linear correspondence between cellular Raman spectra and proteomes and confirm that condition-dependent changes in proteomes can be inferred from the corresponding low-dimensional Raman spectra.

Stoichiometry conservation of proteins in the ISP COG class Since the dimensionality of the proteome data is significantly higher than that of the Raman data, the result above suggests that changes in proteome profiles are constrained in low-dimensional space. The regression matrix B considered above determines how the proteomes relate to the Raman LDA axes. Therefore, analyzing B should provide some insights into constraints on condition-dependent changes in the proteomes (Fig. S1).

The $n \times m$ matrix B is represented as $\mathrm{B} = [\mathbf{b}_0\ \mathbf{b}_1\ \cdots\ \mathbf{b}_{m-1}]$, where the $(k+1)$-th column $\mathbf{b}_k = (b_{1k}\ b_{2k}\ \cdots\ b_{nk})^\top$ $(0 \leq k \leq m-1)$ is the collection of coefficients of all n proteins for the k-th LDA axis (Table 1). In the case of $k = 0$, the coefficients are constant terms. We first asked whether any shared features might exist in the coefficients of B depending on biological functions of corresponding proteins. We then classified the proteins according to functional annotations of Clusters of Orthologous Group (COG) classes (*27–29*) and found that, for many proteins belonging to the "information storage and processing" (ISP) COG class, the

coefficients corresponding to different LDA axes are approximately proportional to the constant terms; i.e., $b_{lk} \approx c_k b_{l0}$, where l is the index of an ISP COG class protein species and c_k is the proportionality constant common to many ISP COG class protein species for the k-th LDA axis (Fig. 3A). The ISP COG class contains various proteins involved in processing genetic information such as translation, transcription, DNA replication, and DNA repair (*20*). Simple calculations show that these proportionality relationships imply that proteins in the ISP COG class conserves their mutual abundance ratios, i.e., stoichiometry, irrespective of environmental conditions (see section 1.3 in (*22*)).

Since this is an implication from the Raman-proteome correspondence, we next examined the stoichiometry conservation only with the proteome data, evaluating the expression levels with Pearson correlation coefficients for all the pairs of the conditions for each COG class (Fig. 3B). For the ISP COG class, the correlation coefficients were close to 1, whereas those for the other COG classes were significantly smaller depending on condition pairs. We also evaluated the coordination of gene expression patterns within each COG class using cosine similarity and obtained consistent results (Fig. S4). Therefore, stoichiometry conservation is stronger in the ISP COG class than in the other COG classes. Remarkably, neither shared transcription factors nor chromosome locations can account for the observed stoichiometry conservation of many protein pairs. Indeed, although the ISP COG class shows highly coordinated expression patterns (Fig. 3C) compared to the non-ISP COG class (Fig. 3D), the gene loci are not chromosomally clustered in either example. Additionally, the similarity/dissimilarity of expression patterns cannot easily be inferred from transcription factor regulation patterns. These results imply multi-level regulation of their abundance.

We consulted other public quantitative proteome data of *Mycobacterium tuberculosis* (*30*), *Mycobacterium bovis* (*30*), and *Saccharomyces cerevisiae* (*31*) under environmental perturbations and consistently found strong stoichiometry conservation of the ISP COG class (Fig. S4).

Furthermore, the same trend was observed for the genotype-dependent expression changes in *E. coli* proteomes (*20*) (Fig. S4).

Identifying stoichiometrically conserved groups Inspired by the existence of a large class of proteins that conserves their stoichiometry, we considered a systematic way to extract stoichiometrically conserved groups (SCGs) without relying on artificial functional classification of COG (Fig. S1). Focusing only on the proteome data, we evaluated stoichiometry conservation for all the pairs of proteins in the proteome by calculating the cosine similarity of expression patterns (i.e., all $\cos\theta_{\mathbf{p}_i\mathbf{p}_j} := (\mathbf{p}_i/\|\mathbf{p}_i\|_2)\cdot(\mathbf{p}_j/\|\mathbf{p}_j\|_2)$ in Fig. 4A and Table 1, where each element of the m-dimensional vector $\mathbf{p}_i$ denotes the expression level of protein species i under one of the m conditions), and extracted groups in each of which the component proteins exhibit coherent expression change patterns by setting a high threshold of cosine similarity (≥ 0.995, Fig. 4B; see section 1.4 in (*22*) for detail).

The largest SCG (SCG 1) included many proteins in the ISP COG class (91 out of 191 SCG 1 members), such as ribosomal proteins and RNA polymerase, and also proteins in the other COG classes (Fig. 4B, Table S4). We call this largest SCG *homeostatic core*, as it constitutes the largest stoichiometry-conserving unit in cells. We found that the abundance of each protein in the homeostatic core (SCG 1) increased approximately linearly with the growth rate in each condition (Fig. 4C). This relationship is reminiscent of the growth law: The total ribosomal contents for translation increase linearly with growth rate (*32–34*). The linear increase in the abundance of each protein in Fig. 4C indicates that the growth law is valid even at the single-gene level for a large class of ribosomal and non-ribosomal proteins in the homeostatic core (Fig. S5) (see section 3.1 in (*22*)).

Though not evenly distributed, the gene loci of the proteins in the homeostatic core are scattered throughout the chromosome (Fig. 4D). Therefore, localization of gene loci to a single

or a small number of operons is not likely a cause of the observed stoichiometry conservation.

The proteins in the second largest SCG (SCG 2) are expressed at high levels in the fast growth conditions, especially in the "LB" condition (Fig. 4C). The SCG 2 includes many proteins in the metabolism COG class (21 out of 26 SCG 2 members) (Table S5), and their abundance increases approximately exponentially with growth rate (Fig. 4C). We also identified other condition-specific small SCGs, such as a group most expressed in the "GlycerolAA" condition (SCG 3) (Table S6), a group mainly expressed in the "Fructose" condition (SCG 4) (Table S7), and a group most expressed in the stationary phase conditions (SCG 5) (Table S8) (Fig. 4C).

Biological relevance of stoichiometry conservation To understand the overall strength of stoichiometry conservation of the proteins in the different SCGs, we calculated the sum of cosine similarity, $d_i = \sum_j \cos\theta_{\mathbf{p}_i\mathbf{p}_j}$, for each protein species i, where $\cos\theta_{\mathbf{p}_i\mathbf{p}_j}$ is cosine similarity between the m-dimensional expression level vectors of protein i and protein j (Fig. 4A), and the sum is taken over all the protein species (see section 1.5 in (*22*)). We refer to d_i as "stoichiometry conservation centrality" (Table 1).

The proteins in the homeostatic core had high centrality scores (Fig. 5A). Therefore, these proteins tend to have more connections with other proteins in terms of stoichiometry conservation. On the other hand, the proteins in the condition-specific SCGs tend to have low centrality scores among all the proteins (Fig. 5A), which suggests that their stoichiometry conservation is localized within each SCG.

The stoichiometry conservation centrality is biologically relevant because it correlates with gene essentiality. Fractions of essential genes almost monotonically decrease with the ranks of centrality score (Fig. 5B and Fig. S6). We also noted that genes with high centrality scores have more orthologs determined by OrthoMCL-DB (*35*) across the three domains of life (Fig.

5C and Fig. S6). Likewise, genes with many orthologs tend to have higher centrality scores (Fig. 5C and Fig. S6). Therefore, the stoichiometry conservation in cells correlates with the evolutionary conservation of proteins.

To determine if the correlation of stoichiometry conservation centrality with gene essentiality and evolutionary conservation is general, we analyzed the transcriptome data from other organisms and found comparable correlations in *S. pombe* (Fig. S6). In addition, we found that fractions of coding genes almost monotonically decreased with ranks of centrality score in the *S. pombe* data (Fig. S6).

We further analyzed two kinds of *Homo sapiens* transcriptome data. One is a human cell atlas, in which expression of both coding and noncoding genes in 15 fetal organs was quantified (*36*), and the other is genome-wide Perturb-seq data (*37*), in which genetically perturbed transcriptomes were measured mainly for coding genes. Our analysis of the human cell atlas data revealed that, while the overall distribution of stoichiometry conservation centrality was broad (Fig. 5D top), the centrality distribution of coding genes was skewed to higher values (Fig. 5D bottom) as observed for the *E. coli* proteome. Fractions of coding genes almost monotonically decreased with ranks of centrality (Fig. 5E) as seen in the *S. pombe* data (Fig. S6). Essentiality of each gene in human cells was quantified by an index called CRISPR score, which measures the fitness cost imposed by CRISPR-based inactivation of the gene (*38*). Genes with lower CRISPR scores are considered more essential. Our analysis revealed that genes with higher stoichiometry conservation centrality scores tend to have lower CRISPR scores, thus more essential (Fig. 5F). Similarly, genes with lower CRISPR scores tend to have higher stoichiometry conservation centrality scores. Furthermore, genes with higher centrality scores have more orthologs across the three domains of life, and vice versa (Fig. 5G). Comparable correlations of stoichiometry conservation with essentiality and evolutionary conservation were also found in the genome-wide Perturb-seq data (Fig. 5H and Fig. 5I). Together, these results suggest

that correlations of stoichiometry conservation centrality with gene essentiality and evolutionary conservation are general and preserved from *E. coli* to human cells regardless of the type of perturbation (see section 1.6 in (*22*) for detail).

Revealing global stoichiometry conservation architecture of the proteomes with csLE Although the previous analysis revealed the biological relevance of stoichiometry conservation centrality, it is a one-dimensional quantity and cannot capture the global architecture of omics profiles. To gain further insights into genome-wide stoichiometry-conserving relationships among genes, we next analyzed the proteomes using a method similar to Laplacian eigenmaps (LE) (Fig. S1) (*39*). We consider a symmetric $n \times n$ matrix A whose (i, j) entry is $\cos\theta_{\mathbf{P}_i\mathbf{P}_j}$ (Fig. 5J, Table 1). The entire proteome structure can be represented using the eigenvectors of normalized A. Major differences of this method from the ordinary LE are that we consider an edge for all node pairs and that we adopt cosine similarity for weighting edges. This method places the proteins with higher cosine similarity closer in the resulting $(m-1)$-dimensional space (see sections 1.5 and 2.1 in (*22*)); we call this linear method cosine similarity LE (csLE).

In this $(m-1)$-dimensional csLE space Ω_{LE}, the stoichiometry conservation centrality of the proteins decreased from center to periphery (Fig. 5K), which confirms that it indeed measures the extent to which each protein is close to the center in the entire stoichiometry conservation architecture. Furthermore, the proteins formed polyhedral distributions with the cluster of the proteins in the homeostatic core at the center and the clusters of the proteins in the other condition-specific SCGs at distinct vertices (Fig. 5L). This distribution is consistent with the fact that the condition-specific SCGs are the components whose expression patterns are distant from the homeostatic core and also between each other.

Representing the proteomes using the Raman LDA axes Given that the analysis of the LDA Raman-proteome regression coefficients B (Fig. 3A) eventually led us to identify the

stoichiometry conservation architecture in the proteome data (Fig. 5), the low-dimensional proteome structure in Ω_{LE} might be related to major changes in cellular Raman spectra in the LDA space and provide insight into the Raman-proteome correspondence. To investigate this, we considered representing the proteomes on the basis of the Raman LDA axes (Fig. S1).

The coefficients in the $n \times m$ regression matrix B must satisfy the proportionality $\frac{b_{ik}}{b_{i0}} = \frac{b_{jk}}{b_{j0}}$ for all k-th LDA axes ($1 \leq k \leq m-1$) for the pair of protein i and protein j that perfectly conserve their stoichiometry, as previously mentioned in the analysis of the ISP COG class (Fig. 3A; see sections 1.3 and 2.1 in (*22*)). Noting this property, we constructed another $(m-1)$-dimensional proteome space Ω_{B}, assigning each protein species i a coordinate $\left(\beta_i^{\mathrm{LDA1}}\ \beta_i^{\mathrm{LDA2}}\ \cdots\ \beta_i^{\mathrm{LDA}m-1}\right)$, where $\beta_i^{\mathrm{LDA}k} := \frac{b_{ik}}{b_{i0}}$ is the normalized coefficient of gene i corresponding to the k-th LDA axis. As in $(m-1)$-dimensional Ω_{LE}, a pair of proteins with strong stoichiometry conservation are expected to position closely in this $(m-1)$-dimensional proteome space Ω_{B}. Note that the proximity of the coordinates $\beta_i^{\mathrm{LDA}k}$ of different proteins i in Ω_{B} is equivalent to the approximate proportionality of different proteins i in Fig. 3A, demonstrated for the ISP COG class using the proportionality constants (normalized coefficients) c_k common to different proteins.

We then found that the distribution of the proteins in Ω_{B} closely resembled the one in Ω_{LE} when visualized using the first few major axes (Fig. 5L and Fig. 6A). This similarity is non-trivial because Ω_{LE} is constructed only from the proteome data, whereas Ω_{B} depends on the $(m-1)$-dimensional Raman LDA space (Fig. 2C–2E).

We remark that each axis of Ω_{B} is directly linked to the corresponding Raman LDA axis. Consequently, the orthants in Ω_{B} where the condition-specific protein species reside agree with those in the Raman LDA space where the cellular Raman spectra under corresponding conditions reside (Fig. S10) (see sections 1.7 and 2.1 in (*22*)). Indeed, we find such orthant agreement between the proteins in the condition-specific SCGs (SCG 2–SCG 5) and the cellular Raman

spectra under the corresponding conditions (Fig. 6B and Fig. 6C). This straightforward correspondence between Ω_{B} and the Raman LDA space allows us to examine the relationship between changes in cellular Raman spectra and omics components' stoichiometry conservation architecture by comparing the two proteome structures in Ω_{B} and Ω_{LE}.

Omics-level interpretation of cellular Raman spectra and a quantitative constraint between expression generality and stoichiometry conservation centrality To understand rigorously what the similarity of the proteome structures in Ω_{B} and Ω_{LE} signifies (Fig. 6C and Fig. 6D), we clarified the mathematical relation between the coordinates of the proteins in these two spaces (Fig. 6E and Fig. S1; see sections 2.1 and 2.2 in (*22*) for detail). We then characterized the two mathematical conditions that must be satisfied simultaneously (Fig. 6E).

The first condition is that major axes of the Raman LDA space and those of the proteome csLE space correspond (Fig. 6E). Consequently, cellular Raman spectra under a condition accompanying the expression of a condition-specific SCG must be significantly different from those under conditions with the expression of other condition-specific SCGs in a manner distinguishable by LDA. Mathematically, this condition is related to the $m \times m$ orthogonal matrix Θ that appears in the equation in Fig. 6E. For the distributions of the proteome components to be similar in the low-dimensional subspaces of Ω_{LE} and Ω_{B}, Θ must be close to the identity matrix with small off-diagonal elements (Fig. 6E). We verified this first condition with the data (Fig. S9; see section 1.8 in (*22*) for detail).

The second condition relates to the proportionality of the n-dimensional vectors $\mathbf{b}_0$ and $\mathbf{b}_0^{\mathrm{est}}$ in Fig. 6E. This proportionality relation can be transformed into another relation that d_i is proportional to $g_i := \|\mathbf{p}_i\|_1/\|\mathbf{p}_i\|_2$, where $\|\mathbf{p}_i\|_1$ and $\|\mathbf{p}_i\|_2$ are the L1 and L2 norms of the expression level m-dimensional vector of protein i across conditions (Fig. 4A and 6E, Table 1).

g_i can be interpreted as *the expression generality score*. When g_i is large, the protein i is

expressed generally across conditions; when g_i is small, this is expressed only under specific conditions (Fig. S8) (see section 1.9 in (*22*)). Therefore, the proportionality between d_i and g_i indicates that the proteins with high stoichiometry conservation centrality must be expressed non-specifically to conditions. We also verified this condition with the data, confirming that it is indeed satisfied (Fig. 7A and Fig. S9).

The spread of the points from the proportionality diagonal line of the *E. coli* proteome data in Fig. 7A was found related to the growth rate under the condition where each protein is expressed the most (see section 2.2 in (*22*) for a detailed analysis on the origin of the deviation). Consequently, one can envisage a growth-rate-dependent expression pattern of each protein on the basis of its relative position in this g_i-d_i plot (Fig. 7B and Fig. 7C). For example, both BamB and YqjD are expressed non-specifically to the conditions with nearly identical expression generality scores. However, BamB is expressed at higher levels under fast growth conditions, whereas YqjP is expressed at higher levels under slow growth conditions due to their relative positions to the proportionality line. Similar growth rate dependence is observed for PaaE and DgoA, but with more prominent condition-specificity because these proteins are characterized by their low expression generality scores. These growth-rate-dependent deviation patterns might hint at a new growth law that governs *the total relative expression changes* of the proteome components (see section 2.2 in (*22*) for detailed discussion).

Generality We also examined the generality of the aforementioned two conditions using the Raman and proteome data of *E. coli* strains with different genotypes (BW25113, MG1655, and NCM3722) under two culture conditions (*20*) and the Raman and transcriptome data of *S. pombe* under 10 culture conditions (*18*). Applying csLE to the omics data, we again found similar omics structures between Ω_{LE} and Ω_{B} when visualized using the first few major axes, with homeostatic cores at the centers and condition-specific SCGs at the vertices (Fig. S11 and

Fig. S12).

Proportionality between stoichiometry conservation centrality and expression generality score was also confirmed in both additional datasets (Fig. S7). We further used publicly available quantitative proteome data of *M. tuberculosis*, *M. bovis*, and *S. cerevisiae* (*30, 31*) to examine this relation and confirmed that the proportionality universally holds (Fig. S7 and Fig. S13). Almost no deviation from the proportionality line existed in the *S. cerevisiae* proteome data measured for the cells in different media but cultured in chemostats with an identical dilution rate (thus, identical growth rate), which is consistent with the result of *E. coli* in which the deviations were related to the growth rate differences.

DISCUSSION

A Raman spectrum obtained from a single cell is a superposition of the spectra of all of its constituent biomolecules. Therefore, cellular Raman spectra potentially contain rich information on essential state differences in targeted cells. The fact that both transcriptomes and proteomes are inferable from cellular Raman spectra, as demonstrated in this and previous (*18*) studies, endorses this speculation. The detailed analyses of the relationship between Raman and omics data have identified functionally relevant constraints on omics changes and provided an interpretation of cellular Raman spectra (Fig. S1). Specifically, it has been revealed that major changes in cellular Raman spectra distinguishable by LDA reflect the changes in omics profiles under the constraints of stoichiometry conservation. This correspondence would help us interpret global changes in cellular Raman spectra by translating them into the differences in omics profiles.

We remark that linearity in our formulation enabled us to find the rigorous connection between the two omics spaces Ω_{B} and Ω_{LE} (Fig. 6E). Unlike the original LE, we adopted cosine similarity as weights of edges between all node pairs to measure expression stoichiometry

conservation of proteins. This modification was indispensable in terms of interpretation; relative proximity of positions in Ω_{LE} reflects the strength of stoichiometry conservation. We also remark that simple principal component analysis (PCA) applied to the normalized *E. coli* proteome data also finds a similar low-dimensional proteome structure (Fig. S6) (see section 1.10 in (*22*)). Therefore, besides interpretability, omics structures in Ω_{LE} might reflect dominant relationships among omics components commonly characterized by several methods of omics representation.

It should be noted that the quantitative analysis of Raman-omics correspondence resulted in the characterization of stoichiometry-conserving architecture in cells (Fig. S1). This shows that besides distinguishing different cellular states or quantifying specific biomolecular species by focusing on spectral peaks, Raman spectra can also characterize the system-level constraints behind changes in global gene expression profiles. While the identified features, such as stoichiometry conservation centrality, expression generality score, and csLE space, can be calculated without Raman data, it is difficult to reach them directly without scrutinizing the Raman-omics correspondence. Furthermore, the definition of expression generality and its relation to stoichiometry conservation centrality were directly derived from the Raman-omics correspondence analysis (Fig. 6E and Fig. 7). Therefore, as a signal reflecting comprehensive molecular profiles in cells, Raman spectra are an important modality for dissecting system-level properties and constraints in cells.

In this study, we mainly analyzed the Raman and proteome data of *E. coli* under 15 different environmental conditions. However, the resulting low-dimensional structures and correspondence of Ω_{LE} and Ω_{B} can change depending on what and how many conditions are included in the analysis. Thus, an intriguing question is how the Raman-proteome correspondence is affected by the conditions used in the analysis. A subsampling analysis focusing on the orthogonal matrix Θ, which represents low-dimensional correspondence precision of Ω_{LE} and Ω_{B}

(Fig. 6E), reveals that correspondence precision tends to increase with an increasing number of conditions (Fig. S14). This result suggests that increasing the number of conditions generally improves the low-dimensional correspondence rather than disrupting it.

Since the proteome data that we referenced (*20*) represent the averaged expression profile of the cells in each condition, we likewise averaged the single-cell Raman data in each condition in the LDA space to determine their correspondence. Once this correspondence is established, it becomes technically feasible to infer the proteomes of individual cells from their Raman spectra. However, verifying the accuracy of the inferred proteome profiles requires quantitative ground truth of single-cell proteomes, which are not yet readily obtainable, especially for bacterial cells. Despite this limitation, future studies may clarify the correspondence at the single-cell level as omics technology advances.

Stoichiometry conservation is plausibly crucial for cellular functions and physiology. For example, the enzymes involved in evolutionarily conserved metabolic pathways conserve their stoichiometry across microorganism species despite their diverse transcriptional and translational rates (*40*). It is suggested that stoichiometry conservation is achieved by optimizing the metabolic flux for fast growth (*41*). Furthermore, a ribosome-targeting antibiotic causes an imbalance of ribosomal proteins and growth arrest in *E. coli*, but the balance is restored alongside growth recovery through physiological adaptation (*42*). These results suggest that disruption of stoichiometric balance among core components could impose significant fitness cost.

It is known that functions, essentiality, and evolutionary conservation of genes can be linked to the topologies of gene networks (*43–48*). However, networks that have been previously analyzed, such as protein-protein interaction networks, depend on known interactions. Therefore, as our understanding of the molecular interactions evolves with new findings, the conclusions may change. Furthermore, analysis of a particular interaction network cannot account for effects of different types of interactions or multilayered regulations affecting each protein species,

thus highlighting only one aspect of the inherently global coordination of molecular compositions in cells. In contrast, the stoichiometry conservation network in this study focuses solely on expression patterns as the net result of interactions and regulations among all types of molecules in cells. Consequently, the stoichiometry conservation networks are not affected by the detailed knowledge of molecular interactions, and naturally reflect the global effects of multilayered interactions behind cellular physiological state changes. Additionally, stoichiometry conservation networks can easily be obtained for non-model organisms, for which detailed molecular interaction information is usually unavailable. Therefore, analysis with the stoichiometry conservation network has several advantages over existing methods from both biological and technical perspectives.

It is intriguing to ask how cells conserve stoichiometry among the components in each SCG. Especially, the homeostatic core (SCG 1) contains many components whose gene loci are scattered throughout the genome. It is known that both transcriptional and translational negative auto-regulation contributes to controlling the stoichiometry of many ribosomal proteins (*49–54*). The genes for the ribosomal proteins are scattered in multiple operons and co-regulated with many other non-ribosomal proteins, such as RNA polymerase subunits, translation initiation/elongation factors, and transmembrane transporters (*55*). Therefore, the stoichiometry-conserving mechanisms established for ribosomes might be partially exploited for the stoichiometry conservation within the homeostatic core.

The existence of condition-specific SCGs and genes with similar expression patterns confirms that adaptation to specific conditions is not necessarily achieved by a small number of functionally relevant genes, but is often accompanied by changes in the expression of many seemingly unrelated genes. Indeed, condition-specific SCGs contain genes with unclear roles in adaptation, including some that are functionally uncharacterized (Tables S5–S8). Therefore, it would be important to investigate whether the coexpression of multiple genes is crucial for

cellular adaptation to a wide range of perturbations while maintaining homeostasis.

The proportionality between stoichiometry conservation centrality and expression generality score suggests that proteins with high stoichiometry conservation centrality govern basal cellular functions required under any conditions. In fact, both essential genes and evolutionarily conserved genes are enriched in the omics fractions with high centrality scores. On the contrary, proteins of low centrality scores might have been acquired in later stages of the evolution and exploited to survive or increase fitness under specific conditions. Such hierarchy in the stoichiometry conservation centrality among core and peripheral processes might promote the adaptability of cells since cells can respond to diverse environments without restructuring a large body of the functional homeostatic core. This architectural principle in omics might underlie the robustness and adaptability of biological cells.

Declaration of interests

YW and KJKK are listed as inventors of patents (JP6993682 and US10,379,052 B2) filed by The University of Tokyo. The other authors declare no competing interests.

Acknowledgments

We thank Matthias Heinemann and Silke Bonsing-Vedelaar for detailed information on the *E. coli* culture conditions; Doeke R. Hekstra, Tetsuya J. Kobayashi, Takafumi Miyamoto, John Russell, and Ian Hunt-Isaak for reading the manuscript and providing critical comments; Kunihiko Kaneko, Chikara Furusawa, Yasushi Okada, and members of the Wakamoto Lab and the Universal Biology Institute for discussion and encouragement. This work was supported by JST CREST Grant Number JPMJCR1927 (Y.W.); JST ERATO Grant Number JPMJER1902 (Y.W.); JSPS KAKENHI Grant Numbers 19J22448 (K.F.K.) and 21K20672 (T.N.).

References

1. C. H. Waddington, *The Strategy of the Genes* (George Allen & Unwin Ltd., 1957).

2. C. H. WADDINGTON, *Nature* **183**, 1654 (1959).

3. R. Goodacre, *et al.*, *Microbiology* **144**, 1157 (1998).

4. Y.-S. Huang, T. Karashima, M. Yamamoto, T. Ogura, H.-o. Hamaguchi, *Journal of Raman Spectroscopy* **35**, 525 (2004).

5. T. Ichimura, *et al.*, *PLoS ONE* **9** (2014).

6. A. Germond, *et al.*, *Communications Biology* **1**, 85 (2018).

7. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proceedings of the National Academy of Sciences* **95**, 14863 (1998).

8. E. Segal, *et al.*, *Nature Genetics* **34**, 166 (2003).

9. S. Bergmann, J. Ihmels, N. Barkai, *Physical Review E* **67**, 031902 (2003).

10. L. Keren, *et al.*, *Molecular Systems Biology* **9**, 701 (2013).

11. C. You, *et al.*, *Nature* **500**, 301 (2013).

12. K. Kaneko, C. Furusawa, T. Yomo, *Physical Review X* **5**, 1 (2015).

13. S. Hui, *et al.*, *Molecular Systems Biology* **11**, 784 (2015).

14. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, *Cell Systems* **2**, 239 (2016).

15. S. Biswas, *et al.*, *Nature Communications* **8** (2017).

16. K. Husain, A. Murugan, *Molecular Biology and Evolution* **37**, 2865 (2020).

17. T. U. Sato, K. Kaneko, *Physical Review Research* **2**, 013197 (2020).

18. K. J. Kobayashi-Kirschvink, *et al.*, *Cell Systems* **7**, 104 (2018).

19. K. J. Kobayashi-Kirschvink, *et al.*, *Nature Biotechnology* pp. 1–9 (2024).

20. A. Schmidt, *et al.*, *Nature Biotechnology* **34**, 104 (2016).

21. T. De Bie, N. Cristianini, R. Rosipal, *Handbook of geometric computing* (Springer, 2005), pp. 129–167.

22. Supplementary materials.

23. L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).

24. R. A. Fisher, *The design of experiments* (Oliver And Boyd; Edinburgh; London, 1937).

25. E. J. Pitman, *Supplement to the Journal of the Royal Statistical Society* **4**, 119 (1937).

26. B. Phipson, G. K. Smyth, *Statistical Applications in Genetics and Molecular Biology* **9** (2010).

27. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).

28. R. L. Tatusov, *et al.*, *BMC Bioinformatics* **4**, 1 (2003).

29. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, *Nucleic Acids Research* **43**, D261 (2015).

30. O. T. Schubert, *et al.*, *Cell Host & Microbe* **18**, 96 (2015).

31. P.-J. Lahtvee, *et al.*, *Cell Systems* **4**, 495 (2017).

32. F. C. Neidhardt, B. Magasanik, *Biochimica et Biophysica Acta* **42**, 99 (1960).

33. M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, T. Hwa, *Science* **330**, 1099 (2010).

34. H. Bremer, P. P. Dennis, *EcoSal Plus* **3** (2008).

35. F. Chen, A. J. Mackey, C. J. Stoeckert Jr, D. S. Roos, *Nucleic Acids Research* **34**, D363 (2006).

36. J. Cao, *et al.*, *Science* **370**, eaba7721 (2020).

37. J. M. Replogle, *et al.*, *Cell* **185**, 2559 (2022).

38. T. Wang, *et al.*, *Science* **350**, 1096 (2015).

39. M. Belkin, P. Niyogi, *Neural Computation* **15**, 1373 (2003).

40. J.-B. Lalanne, *et al.*, *Cell* **173**, 749 (2018).

41. J.-B. Lalanne, G.-W. Li, *Elife* **10**, e69222 (2021).

42. Y. Koganezawa, M. Umetani, M. Sato, Y. Wakamoto, *Elife* **11**, e74486 (2022).

43. H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai, *Nature* **411**, 41 (2001).

44. X. He, J. Zhang, *PLoS genetics* **2**, e88 (2006).

45. H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, *PLoS computational biology* **3**, e59 (2007).

46. H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, M. W. Feldman, *Science* **296**, 750 (2002).

47. S. Wuchty, Z. N. Oltvai, A.-L. Barabási, *Nature genetics* **35**, 176 (2003).

48. X. Li, W. Li, M. Zeng, R. Zheng, M. Li, *Briefings in bioinformatics* **21**, 566 (2020).

49. M. Nomura, J. L. Yates, D. Dean, L. E. Post, *Proceedings of the National Academy of Sciences* **77**, 7084 (1980).

50. D. Dean, J. L. Yates, M. Nomura, *Nature* **289**, 89 (1981).

51. M. Kaczanowska, M. Rydén-Aulin, *Microbiology and Molecular Biology Reviews* **71**, 477 (2007).

52. C. Portier, M. Grunberg-Manago, *Translational Regulation of Gene Expression 2* (Springer, 1993), pp. 23–47.

53. L. V. Aseev, A. A. Levandovskaya, L. S. Tchufistova, N. V. Scaptsova, I. V. Boni, *RNA* **14**, 1882 (2008).

54. B. Roy, *et al.*, *Communications Biology* **3**, 1 (2020).

55. I. M. Keseler, *et al.*, *Nucleic Acids Research* **45**, D543 (2017).

56. K. L. Howe, *et al.*, *Nucleic Acids Research* **48**, D689 (2020).

Figure legends

Fig. 1. Cellular physiological state differences detected by Raman spectral global patterns and gene expression profiles. (A) Condition-dependent cellular Raman spectral patterns. Raman spectra obtained from cells reflect their molecular profiles. Therefore, systematic differences in global spectral patterns may indicate their physiological states. A Raman spectrum from each cell can be represented as a vector and a point in a high-dimensional Raman space. If condition-dependent differences exist in the spectral patterns, appropriate dimensional reduction methods allow us to classify the spectra and detect cellular physiological states in a low-dimensional space. **(B)** Condition-dependent gene expression profiles. Global gene expression profiles (proteomes and transcriptomes) are also dependent on conditions. For each gene, we can consider a high-dimensional vector whose elements represent expression levels under different conditions. It has been suggested that these expression-level vectors are constrained to some low-dimensional manifolds (*7–17*). This study characterizes the statistical correspondence between dimension-reduced Raman spectral patterns and gene expression profiles. Analyzing the correspondence, we reveal a stoichiometry conservation principle that constrains gene expression profiles to low-dimensional manifolds.

Fig. 2. Estimation of proteomes from Raman spectra. (A) The experimental design. We cultured *E. coli* cells under 15 different conditions and measured single cells' Raman spectra. We then examined the correspondence between the measured Raman spectra and the absolute quantitative proteome data reported by Schmidt *et al.* (*20*). **(B)** Representative Raman spectra from single cells, one from the "Glucose" condition, and the other from the "LB" condition. The fingerprint region and representative peaks are annotated. **(C to E)** Cellular Raman spectra in LDA space. The dimensionality of the spectra is reduced to $14\ (= 15 - 1)$. Each point

represents a spectrum from a single cell, and each ellipse shows the 95% concentration ellipse for each condition. Their projection to the LDA1-LDA2 plane (C), the LDA1-LDA3 plane (D), and the LDA1-LDA4 plane (E) are shown. **(F)** Visualization of the 14 dimensional LDA space embedded in two dimensional space with t-SNE. **(G)** The scheme of leave-one-out cross-validation. The Raman and proteome data of one condition (here j) are excluded, and the matrix B is estimated using the data of the rest of the conditions as $\mathrm{B}^{\mathrm{est}}_{-j}$. The proteome data under the condition j is estimated from the Raman data $\hat{\mathbf{r}}_j$ with $\mathrm{B}^{\mathrm{est}}_{-j}$ and compared with the actual data to calculate estimation errors. **(H)** Comparison of measured and estimated proteome data. The plot for the "Glucose" condition is shown as an example. Each dot corresponds to one protein species. The straight line indicates $x = y$. Proteins with negative estimated values are not shown.

Fig. 3. A stoichiometrically conserved protein group identified by an analysis of the Raman-proteome coefficient matrix. **(A)** Scatterplots of Raman-proteome transformation coefficients. The horizontal axes are constant terms ($\mathbf{b}_0$) in all the plots. The vertical axis is coefficients for LDA1 ($\mathbf{b}_1$), LDA2 ($\mathbf{b}_2$), LDA3 ($\mathbf{b}_3$), or LDA4 ($\mathbf{b}_4$) in each plot. The proteins in the ISP COG class are indicated in yellow. Yellow solid straight lines are least squares regression lines passing through the origins for the ISP proteins. Insets are enlarged views of area around the origins. In this figure, we used the average of $\mathrm{B}^{\mathrm{est}}_{-i}$ as an estimate of B. **(B)** Similarity of expression patterns between culture conditions for each COG class. We divided the proteome into COG classes (*28, 29*), and calculated Pearson's correlation coefficient of expression patterns for all the combinations of culture conditions. Since the data are from 15 conditions, there are 105 ($= 15{\cdot}14/(2{\cdot}1)$) points for each COG class in the graph. The box-and-whisker plots summarize the distributions of the points. The lines inside the boxes denote the medians, the top and bottom edges of the boxes do the 25th percentiles and 75th percentiles, respectively. The numbers

of protein species are 376 for the Cellular Processes and Siginaling COG class, 354 for the ISP COG class, and 840 for the Metabolism COG class. See Fig. S4 for the evaluation with Pearson correlation coefficient of log abundances and with cosine similarity. Fig. S4 also contains figures directly showing expression-level changes of different protein species across conditions for each COG class. **(C)** Examples of stoichiometry-conserving proteins in the ISP COG class. The horizontal axis represents the abundance of RplF under 15 conditions and the vertical axis represents those of several ISP COG class proteins. These proteins are also contained in the *homeostatic core* defined later (see Fig. 4). The solid straight lines are linear regression lines with intercept of zero. **(D)** Examples of abundance ratios of non-ISP COG class proteins. The horizontal axis represents the abundance of RplF under 15 conditions and the vertical axis represents those of compared non-ISP COG class proteins. Crp belongs to the Cellular Processes and Signaling COG class; the other proteins belong to the Metabolism COG class. In both (C) and (D), we selected the proteins expressed from distant loci on the chromosome. All sigma factors participating in the regulation of the proteins examined in (C) and (D) are listed on the right of the gene name legends. All transcription factors known to regulate multiple genes listed here are shown in the right diagrams. Arrows show activation; bars represent inhibition; and squares indicate that a transcription factor activates or inhibits depending on other factors. The information on gene regulation and functions was obtained from EcoCyc (*55*) in August 2022. The error bars are standard errors calculated by using the data of (*20*). The inset shows the positions of the genes on the *E. coli* chromosome determined based on ASM75055v1.46 (*56*). No genes are in the same operon.

Fig. 4. Extracting SCGs from proteome data. (A) Quantifying stoichiometry conservation by cosine similarity. We consider an m-dimensional expression vector for each protein species whose elements represent its abundance under different conditions. The cosine similarity be-

tween the m-dimensional expression vectors of two protein species becomes nearly 1 when they conserve mutual stoichiometry strongly across conditions, whereas lower than 1 when their expression patterns are incoherent. **(B)** Extracted SCGs. We extracted proteins with high cosine similarity relationships. Each node represents a protein species. An edge connecting two nodes represents that the expression patterns of the two connected protein species have high cosine similarity exceeding a threshold of 0.995. Proteins that have no edge with the other proteins are not shown. The largest and the second largest protein groups, which we refer to as SCG 1 and SCG 2, respectively, are indicated by shaded polygons. **(C)** Expression patterns of the extracted SCGs. The horizontal and vertical axes represent growth rate and protein abundance, respectively. Line-connected points represent expression-level changes of different protein species across conditions. SCG 1 (homeostatic core) is shown in two ways: the left panel with a linear-scaled vertical axis and the right panel with a log-scaled vertical axis. The inset for SCG 2 shows the total abundances of SCG 2 proteins with a log-scaled vertical axis. Error bars are standard errors. **(D)** The gene loci of the homeostatic core (SCG 1) proteins on the chromosome. Magenta dots are nodes (genes), and gray lines are edges (high cosine similarity relationships). We determined the gene loci based on ASM75055v1.46 (*56*).

Fig. 5. A proteome structure characterized by global stoichiometry conservation relationships. (A) Distributions of stoichiometry conservation centrality values for all the proteins (gray), the homeostatic core (SCG 1) proteins (magenta), and the proteins belonging to the other SCGs (cyan). **(B)** Correlation between stoichiometry conservation centrality and gene essentiality. The proportion of essential genes within each class of stoichiometry conservation ranking is shown. The list of essential genes was downloaded from EcoCyc (*55*). **(C)** Correlation between stoichiometry conservation and evolutionary conservation. The strength of evolutionary conservation of each protein species was estimated by the number of orthologs found in the OrthoMCL

species (*35*). The genes with more orthologs tend to have higher stoichiometry conservation centrality ($p = 3.42 \times 10^{-14}$ by one-sided Brunner–Munzel test between the top 25% and the bottom 25% fractions of ortholog number ranking). Likewise, the genes with higher stoichiometry conservation centrality scores tend to have more orthologs ($p = 8.44 \times 10^{-12}$ by one-sided Brunner–Munzel test, Top 25%–Bottom 25% comparison; p-values in the captions for (F) to (I) were evaluated with the same statistical test scheme). **(D) to (G)** Stoichiometry conservation analyses of human cell atlas transcriptome data of fetal 15 organs (*36*). The top gray histogram in (D) shows the distribution of stoichiometry conservation centrality values for all genes. The bottom histograms in (D) show the distribution for coding genes (yellow) and that for the other genes (cyan). (E) shows a correlation between the ratio of coding genes and stoichiometry conservation centrality calculated from the human cell atlas data. (F) shows a correlation between gene essentiality and stoichiometry conservation centrality calculated from the human cell atlas data. The essentiality of each human gene was quantified by CRISPR score, which is the fitness cost imposed by CRISPR-based inactivation of the gene in KBM7 chronic myelogenous leukemia cells (*35*). Genes with lower CRISPR score are regarded as more essential. The fraction with low CRISPR score (i.e. high essentiality fraction) tends to have higher stoichiometry conservation centrality ($p < 10^{-15}$). The fraction with high centrality score tends to be more essential ($p < 10^{-15}$). (G) shows a correlation between evolutionary conservation and stoichiometry conservation centrality based on the human cell atlas data. The gene fraction with many orthologs tends to have higher stoichiometry conservation centrality ($p < 10^{-15}$). The gene fraction with high centrality score tends to have more orthologs ($p < 10^{-15}$). **(H) and (I)** Stoichiometry conservation analyses of genome-wide Perturb-seq data (*37*). (H) shows a correlation between stoichiometry conservation centrality calculated from the Perturb-seq data and gene essentiality. The essentiality of each gene was quantified by the CRISPR score as in (F). The gene fraction with low CRISPR score (i.e. high essentiality fraction) tends to have higher

stoichiometry conservation centrality ($p < 10^{-15}$). The gene fraction with high centrality score tends to be more essential ($p < 10^{-15}$). (I) shows a correlation between stoichiometry conservation based on the Perturb-seq data and evolutionary conservation of genes. The gene fraction with many orthologs tends to have higher stoichiometry conservation centrality ($p < 10^{-15}$). The gene fraction with high centrality score tends to have more orthologs ($p < 10^{-15}$). **(J)** Representation of the proteomes as a graph. A node corresponds to a protein species, and the weight of an edge is taken as the cosine similarity between the m-dimensional expression vectors of the two connected protein species. The $n \times n$ matrix A can specify the whole graph. Note that the diagonal elements of A are ones, which were introduced just for simplicity. **(K)** Cosine similarity LE (csLE) structure in a three-dimensional space. Each dot represents a different protein species and is color-coded on the basis of its stoichiometry conservation centrality value. We selected the axes considering the structural similarity to the Raman-based proteome structure in Ω_{B} (see Fig. 6). **(L)** The csLE structure in a three-dimensional space. The views from two different angles are shown. Each gray dot represents a different protein species. The proteins belonging to each SCG are indicated with distinct markers. Colors of the two-dimensional histograms in (C), (F), (G), (H), and (I) represent the height of each bar.

Fig. 6. Raman-based proteome structure and its similarity to stoichiometry-based proteome structure. (A) Proteome structure determined by Raman-proteome coefficients visualized in a three-dimensional space. The views from two different angles are shown. Each gray dot represents a protein species. The proteins belonging to each SCG are indicated with distinct markers. We note that SCGs are defined without referring to Raman data (Fig. 4). **(B–D)** Similarity among the distribution of LDA Raman spectra (B), the proteome structure determined by Raman-proteome coefficients (C), and the proteome structure determined by stoichiometry conservation (D). **(E)** Mathematical relation between the coordinates of the proteins in Ω_{B} (C)

and Ω_{LE} (D). The two conditions, one with Θ (magenta), and the other between $\mathbf{b}_0$ and $\mathbf{b}_0^{\mathrm{est}}$ (cyan), must hold for the similarity between the two proteome structures (yellow), as described in the gray box. $\overset{\star}{\propto}$ denotes column-wise proportionality.

Fig. 7. Proportionality between stoichiometry conservation centrality and expression generality. (A) Relationships between stoichiometry conservation centrality (d_i) and expression generality (g_i). Each gray dot represent a protein species. The proteins belonging to each SCG are indicated with distinct markers. The dashed lines are $y = n, x = 1, \sqrt{m}$ $(n = 2058, m = 15)$. The solid lines represent $y = \left\{\left(\sum_{j=1}^{n} d_j\right)/m\right\}^{1/2} x$ (see section 2.2 in (*22*)). The deviation of a point from the solid line is related to the growth rate under the condition where each protein is expressed the most. **(B)** The same plot as (A) in black and white. Overlaid red circles indicate proteins featured in (C). **(C)** Expression patterns of the proteins indicated by red circles in (B) across conditions. The condition differences are shown by the growth rate differences on the horizontal axes. The arrangement of the plots for the proteins corresponds to their relative positions in (B).

Tables

Table 1: **List of scalars, vectors and matrices in the main text.**
Scalars, vectors, and matrices in the main text are listed with their sizes and descriptions. m is the number of conditions, and n is the number of protein species. ($m = 15$ and $n = 2058$ in the main text.)

	Size (#columns $\times$ #rows)	**Description**
$\hat{\mathbf{r}}_j$ $(j = 1, \ldots, m)$	$(m-1) \times 1$ (vector)	mean LDA Raman profile of single cells under condition j
$\hat{\mathbf{p}}_j$ $(j = 1, \ldots, m)$	$n \times 1$ (vector)	proteome profile of cell population under condition j
B $= [\mathbf{b}_0 \ \cdots \ \mathbf{b}_{m-1}]$ $= (b_{ik})_{1 \leq i \leq n, 0 \leq k \leq m-1}$	$n \times m$	set of condition-independent coefficients that linearly connect $\hat{\mathbf{r}}_j$ and $\hat{\mathbf{p}}_j$ for all conditions j (Equation (1))
$\mathbf{p}_i$ $(i = 1, \ldots, n)$	$m \times 1$ (vector)	expression levels of protein species i across m conditions
$\cos\theta_{\mathbf{p}_i\mathbf{p}_j}$ $= (\mathbf{p}_i/\lVert\mathbf{p}_i\rVert_2) \cdot (\mathbf{p}_j/\lVert\mathbf{p}_j\rVert_2)$ $(i, j = 1, \ldots, n)$	1×1 (scalar)	stoichiometry (abundance ratio) conservation strength between two protein species i and j (Fig. 4A)
$\mathrm{A} = ((\cos\theta_{\mathbf{p}_i\mathbf{p}_j})_{ij})_{1 \leq i,j \leq n}$	$n \times n$	set of stoichiometry conservation strengths between all pairs of protein species (Fig. 5J)
$d_i = \Sigma_{j=1}^{n} \cos\theta_{\mathbf{p}_i\mathbf{p}_j}$ $(i = 1, \ldots, n)$	1×1 (scalar)	stoichiometry conservation centrality of protein species i
$g_i = \lVert\mathbf{p}_i\rVert_1/\lVert\mathbf{p}_i\rVert_2$ $(i = 1, \ldots, n)$	1×1 (scalar)	expression generality of protein species i

Fig. 1

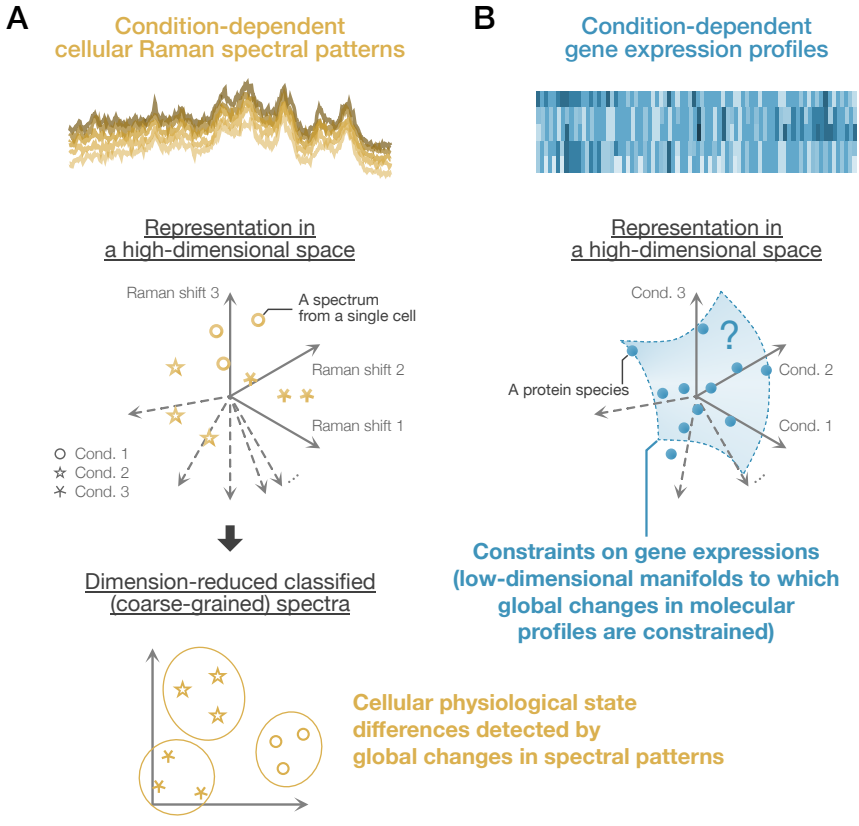

A
Condition-dependent
cellular Raman spectral patterns
Representation in
a high-dimensional space
Raman shift 3
A spectrum
from a single cell
Raman shift 2
Raman shift 1
Cond. 1
Cond. 2
Cond. 3
Dimension-reduced classified
(coarse-grained) spectra
B
Condition-dependent
gene expression profiles
Representation in
a high-dimensional space
Cond. 3
?
Cond. 2
A protein species
Cond. 1
Constraints on gene expressions
(low-dimensional manifolds to which
global changes in molecular
profiles are constrained)
Cellular physiological state
differences detected by
global changes in spectral patterns

Fig. 2

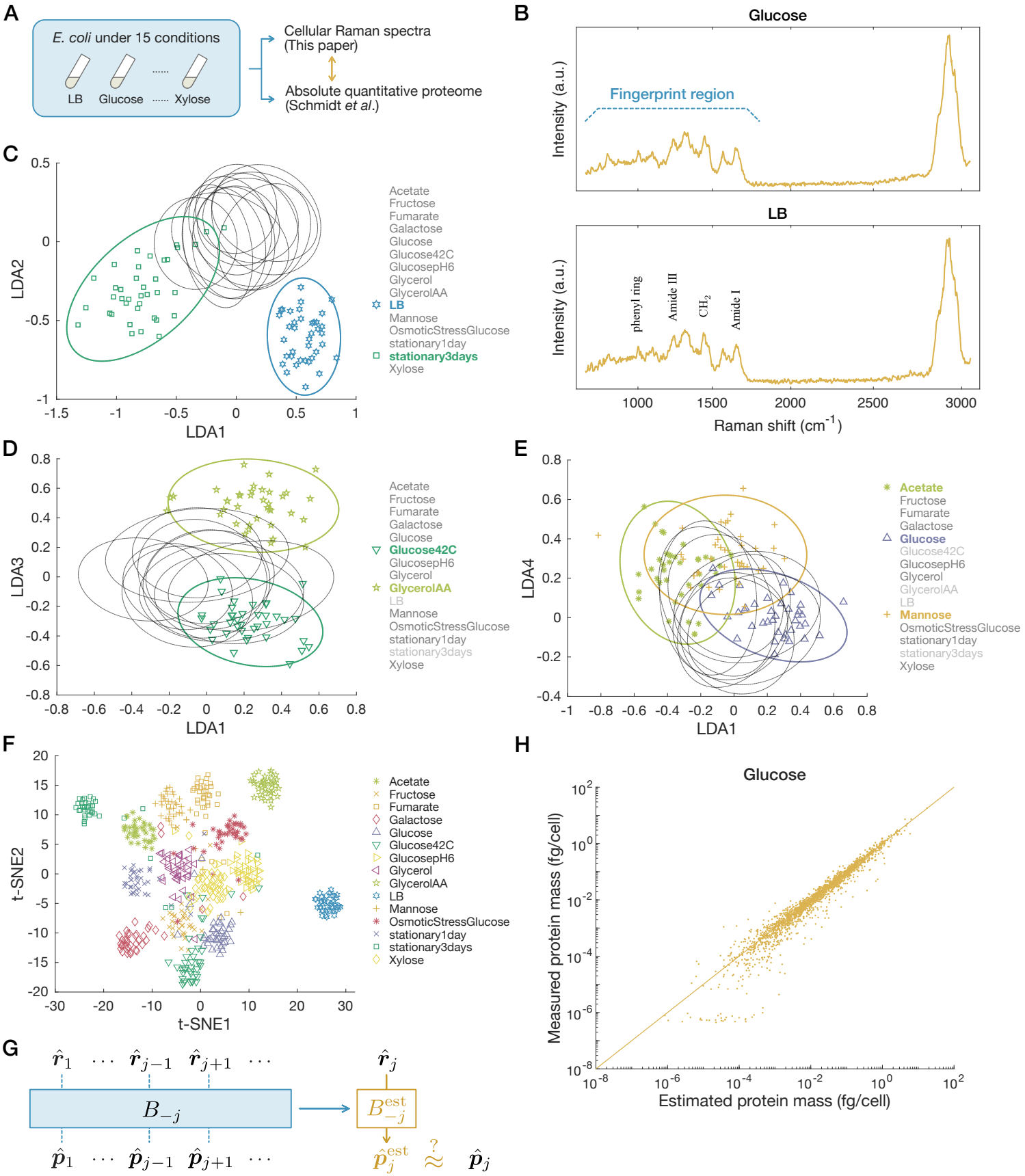

A
E. coli under 15 conditions
LB
Glucose
Xylose
Cellular Raman spectra
(This paper)
Absolute quantitative proteome
(Schmidt et al.)
B
Glucose
Fingerprint region
Intensity (a.u.)
LB
phenyl ring
Amide III
CH2
Amide I
Raman shift (cm-1)
C
LDA1
LDA2
Acetate
Fructose
Fumarate
Galactose
Glucose
Glucose42C
GlucosepH6
Glycerol
GlycerolAA
LB
Mannose
OsmoticStressGlucose
stationary1day
stationary3days
Xylose
D
LDA1
LDA3
E
LDA1
LDA4
F
t-SNE1
t-SNE2
G
H
Glucose
Measured protein mass (fg/cell)
Estimated protein mass (fg/cell)

Fig. 3

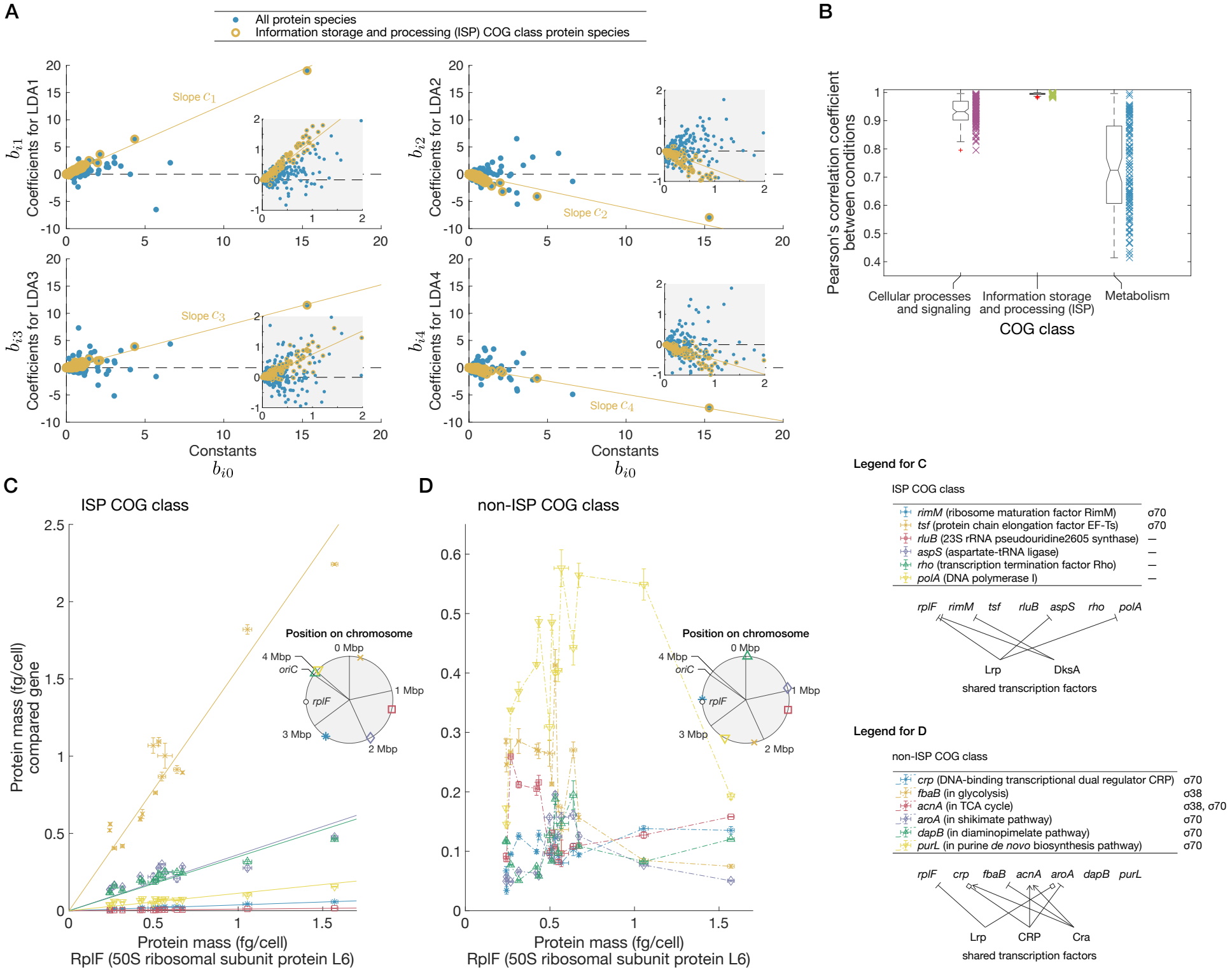

A
All protein species
Information storage and processing (ISP) COG class protein species
Coefficients for LDA1
Coefficients for LDA2
Coefficients for LDA3
Coefficients for LDA4
Slope c_1
Slope c_2
Slope c_3
Slope c_4
Constants
b_{i0}
B
Pearson's correlation coefficient between conditions
Cellular processes and signaling
Information storage and processing (ISP)
Metabolism
COG class
C
ISP COG class
Protein mass (fg/cell) compared gene
Protein mass (fg/cell)
RplF (50S ribosomal subunit protein L6)
Position on chromosome
0 Mbp
1 Mbp
2 Mbp
3 Mbp
4 Mbp
oriC
rplF
D
non-ISP COG class
Legend for C
ISP COG class
rimM (ribosome maturation factor RimM) σ70
tsf (protein chain elongation factor EF-Ts) σ70
rluB (23S rRNA pseudouridine2605 synthase) —
aspS (aspartate-tRNA ligase) —
rho (transcription termination factor Rho) —
polA (DNA polymerase I) —
rplF rimM tsf rluB aspS rho polA
Lrp DksA
shared transcription factors
Legend for D
non-ISP COG class
crp (DNA-binding transcriptional dual regulator CRP) σ70
fbaB (in glycolysis) σ38
acnA (in TCA cycle) σ38, σ70
aroA (in shikimate pathway) σ70
dapB (in diaminopimelate pathway) σ70
purL (in purine de novo biosynthesis pathway) σ70
rplF crp fbaB acnA aroA dapB purL
Lrp CRP Cra
shared transcription factors

Fig. 4

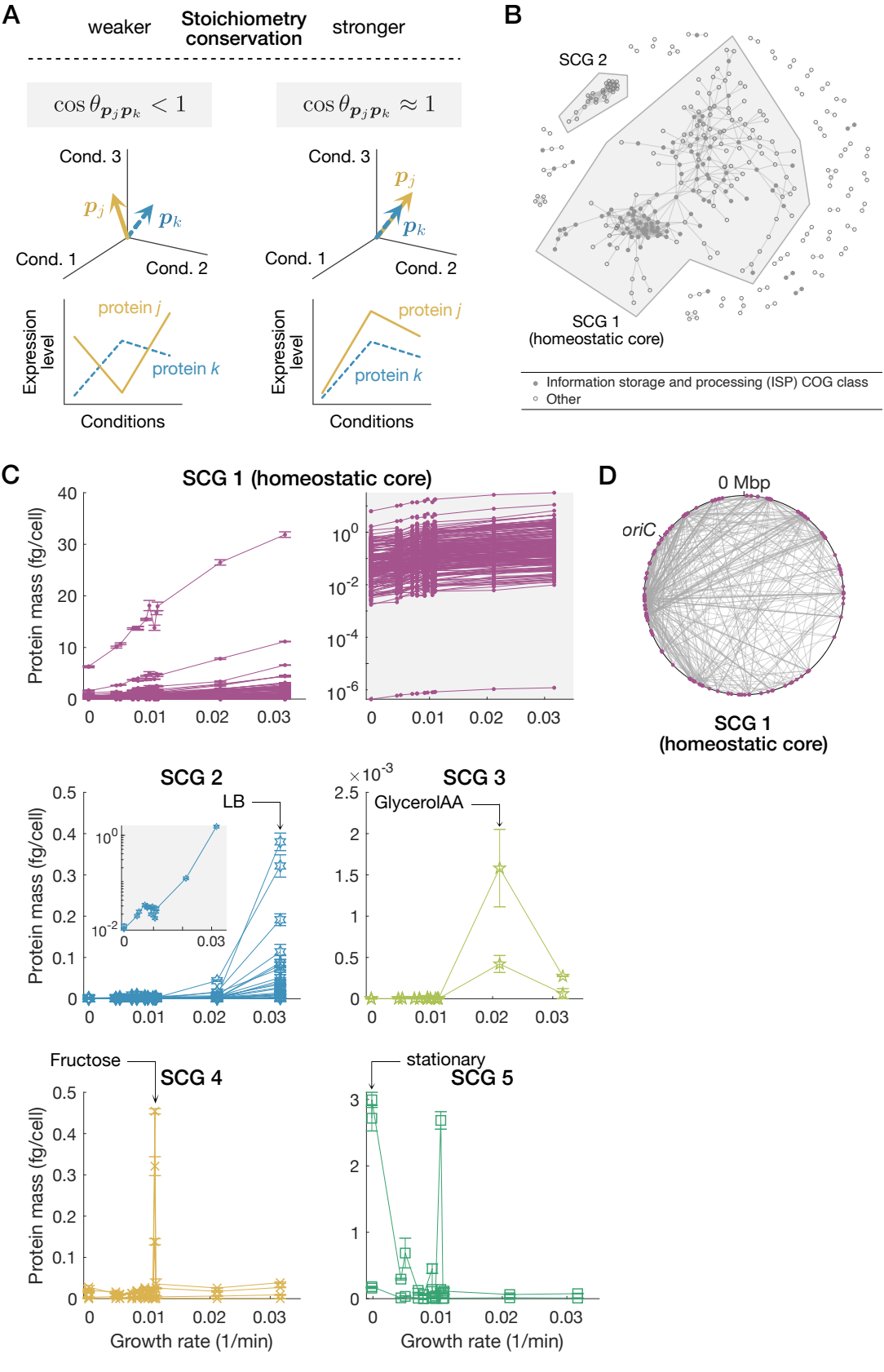

A
weaker
Stoichiometry conservation
stronger
$\cos\theta_{\boldsymbol{p}_j\boldsymbol{p}_k} < 1$
$\cos\theta_{\boldsymbol{p}_j\boldsymbol{p}_k} \approx 1$
Cond. 3
Cond. 1
Cond. 2
$\boldsymbol{p}_j$
$\boldsymbol{p}_k$
Expression level
protein j
protein k
Conditions
B
SCG 2
SCG 1
(homeostatic core)
Information storage and processing (ISP) COG class
Other
C
SCG 1 (homeostatic core)
Protein mass (fg/cell)
SCG 2
LB
SCG 3
GlycerolAA
SCG 4
Fructose
SCG 5
stationary
Growth rate (1/min)
D
0 Mbp
oriC
SCG 1
(homeostatic core)

Fig. 5

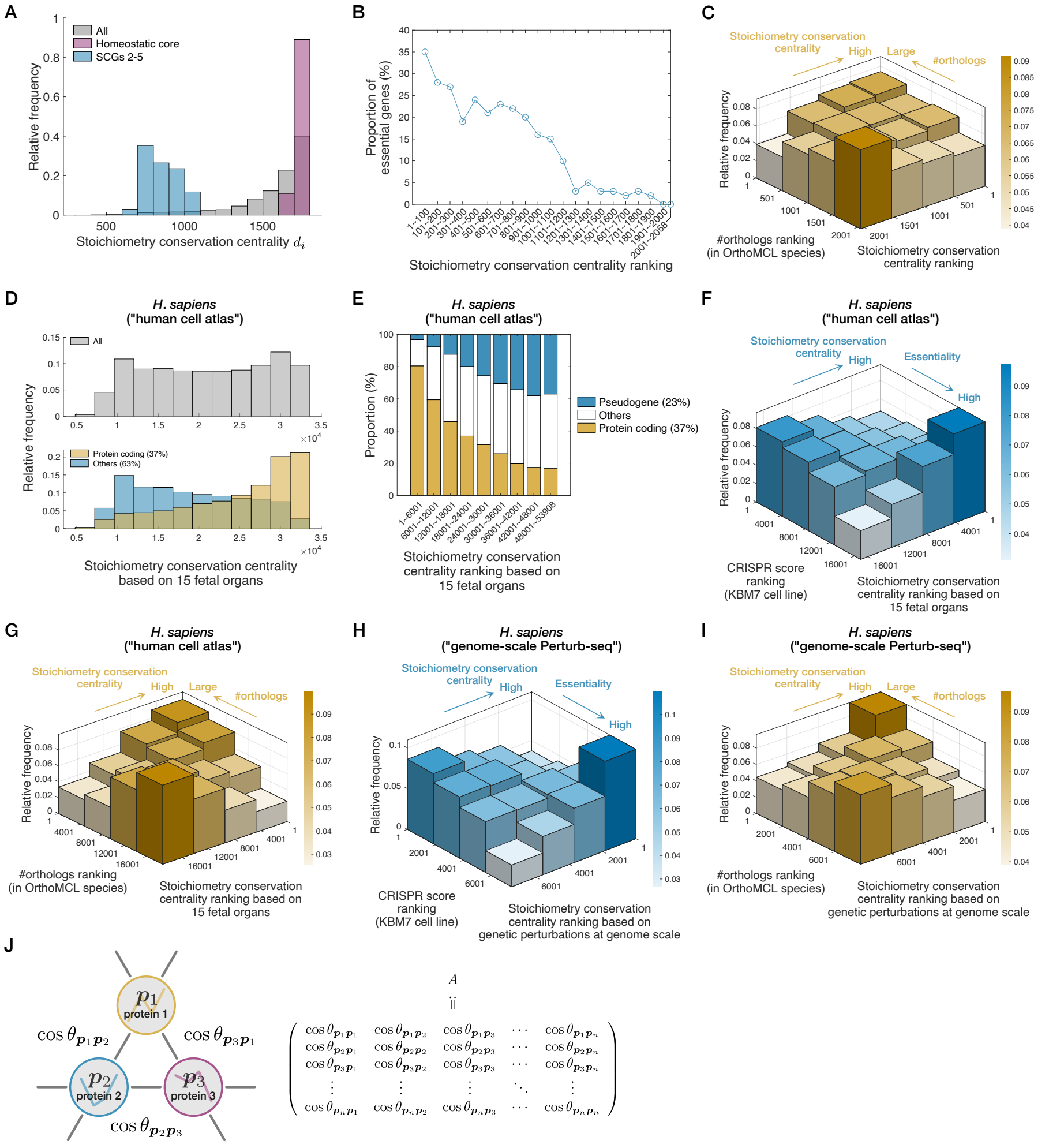
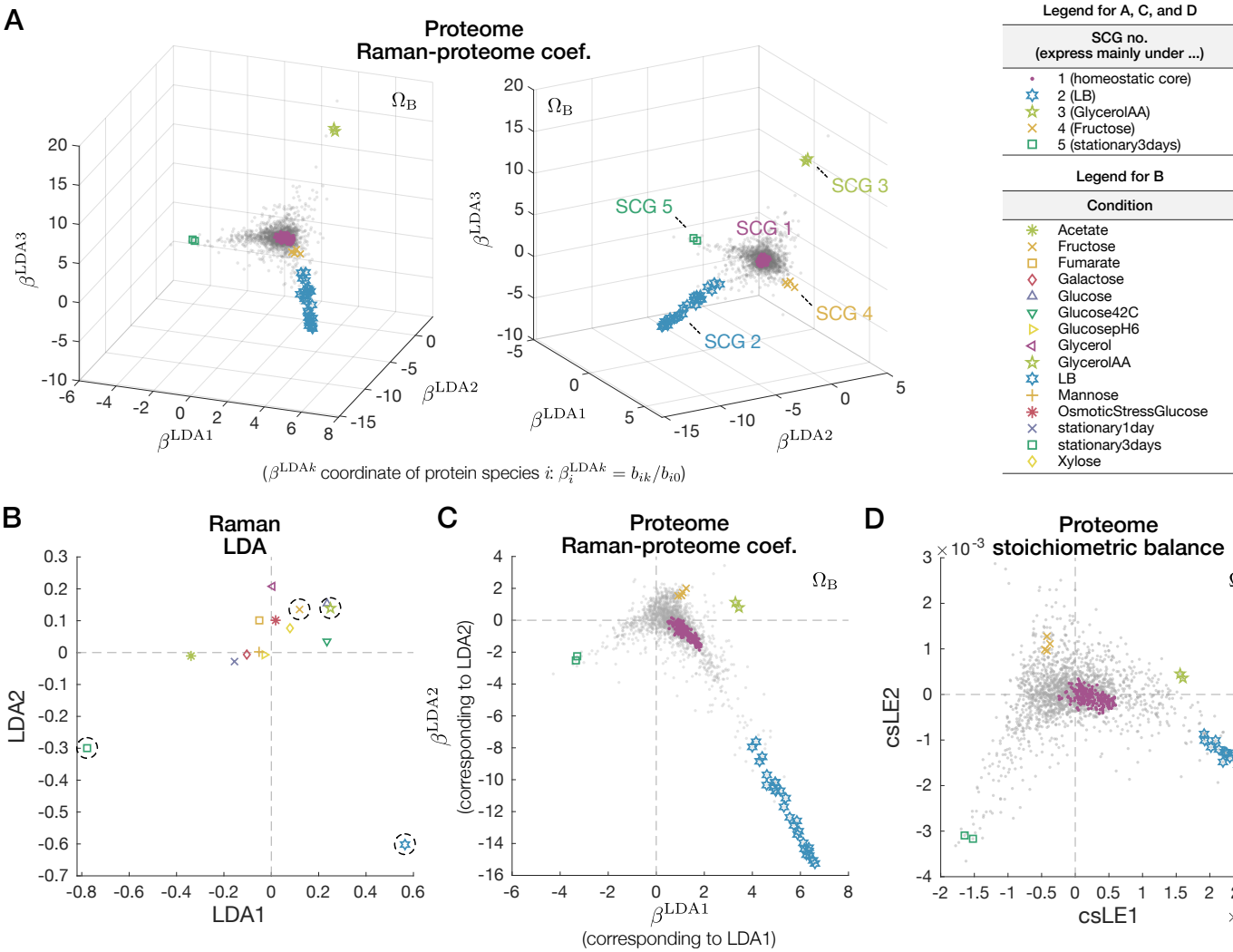

A
All
Homeostatic core
SCGs 2-5
Relative frequency
Stoichiometry conservation centrality d_i
B
Proportion of essential genes (%)
Stoichiometry conservation centrality ranking
C
Stoichiometry conservation centrality
High
Large
#orthologs
Relative frequency
#orthologs ranking (in OrthoMCL species)
Stoichiometry conservation centrality ranking
D
H. sapiens ("human cell atlas")
All
Protein coding (37%)
Others (63%)
Relative frequency
Stoichiometry conservation centrality based on 15 fetal organs
E
H. sapiens ("human cell atlas")
Proportion (%)
Pseudogene (23%)
Others
Protein coding (37%)
Stoichiometry conservation centrality ranking based on 15 fetal organs
F
H. sapiens ("human cell atlas")
Stoichiometry conservation centrality
High
Essentiality
High
Relative frequency
CRISPR score ranking (KBM7 cell line)
Stoichiometry conservation centrality ranking based on 15 fetal organs
G
H. sapiens ("human cell atlas")
Stoichiometry conservation centrality
High
Large
#orthologs
Relative frequency
#orthologs ranking (in OrthoMCL species)
Stoichiometry conservation centrality ranking based on 15 fetal organs
H
H. sapiens ("genome-scale Perturb-seq")
Stoichiometry conservation centrality
High
Essentiality
High
Relative frequency
CRISPR score ranking (KBM7 cell line)
Stoichiometry conservation centrality ranking based on genetic perturbations at genome scale
I
H. sapiens ("genome-scale Perturb-seq")
Stoichiometry conservation centrality
High
Large
#orthologs
Relative frequency
#orthologs ranking (in OrthoMCL species)
Stoichiometry conservation centrality ranking based on genetic perturbations at genome scale
J
p_1 protein 1
p_2 protein 2
p_3 protein 3
$\cos\theta_{p_1p_2}$
$\cos\theta_{p_3p_1}$
$\cos\theta_{p_2p_3}$
A
$\cos\theta_{p_1p_1}$ $\cos\theta_{p_1p_2}$ $\cos\theta_{p_1p_3}$ $\cdots$ $\cos\theta_{p_1p_n}$
$\cos\theta_{p_2p_1}$ $\cos\theta_{p_2p_2}$ $\cos\theta_{p_2p_3}$ $\cdots$ $\cos\theta_{p_2p_n}$
$\cos\theta_{p_3p_1}$ $\cos\theta_{p_3p_2}$ $\cos\theta_{p_3p_3}$ $\cdots$ $\cos\theta_{p_3p_n}$
$\cos\theta_{p_np_1}$ $\cos\theta_{p_np_2}$ $\cos\theta_{p_np_3}$ $\cdots$ $\cos\theta_{p_np_n}$
K
Proteome stoichiometric balance
Ω_{LE}
csLE1
csLE2
csLE4
Stoichiometry conservation centrality d_i
L
Proteome stoichiometric balance
Ω_{LE}
csLE1
csLE2
csLE4
SCG 1
SCG 2
SCG 3
SCG 4
SCG 5

Fig. 6



E

LDA Raman-based matrix which contains protein coordinates in Ω_B

Stoichiometry-based matrix which contains protein coordinates in Ω_{LE}

$$\sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right) B_E^{\mathrm{norm}} \Sigma_{R_E}^{\mathrm{norm}} = \mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right) B_E^{\mathrm{est,norm}} \Sigma_{\mathrm{LE}} \Theta$$

... Supplementary materials equation (2.138)

Vector which contains constant terms for Raman-proteome transformation

Vector which contains estimated constant terms for Raman-proteome transformation

Orthogonal transformation (e.g. rotation)

$B_E^{\mathrm{norm}} \overset{\star}{\propto} B_E^{\mathrm{est,norm}} \quad \Leftrightarrow$

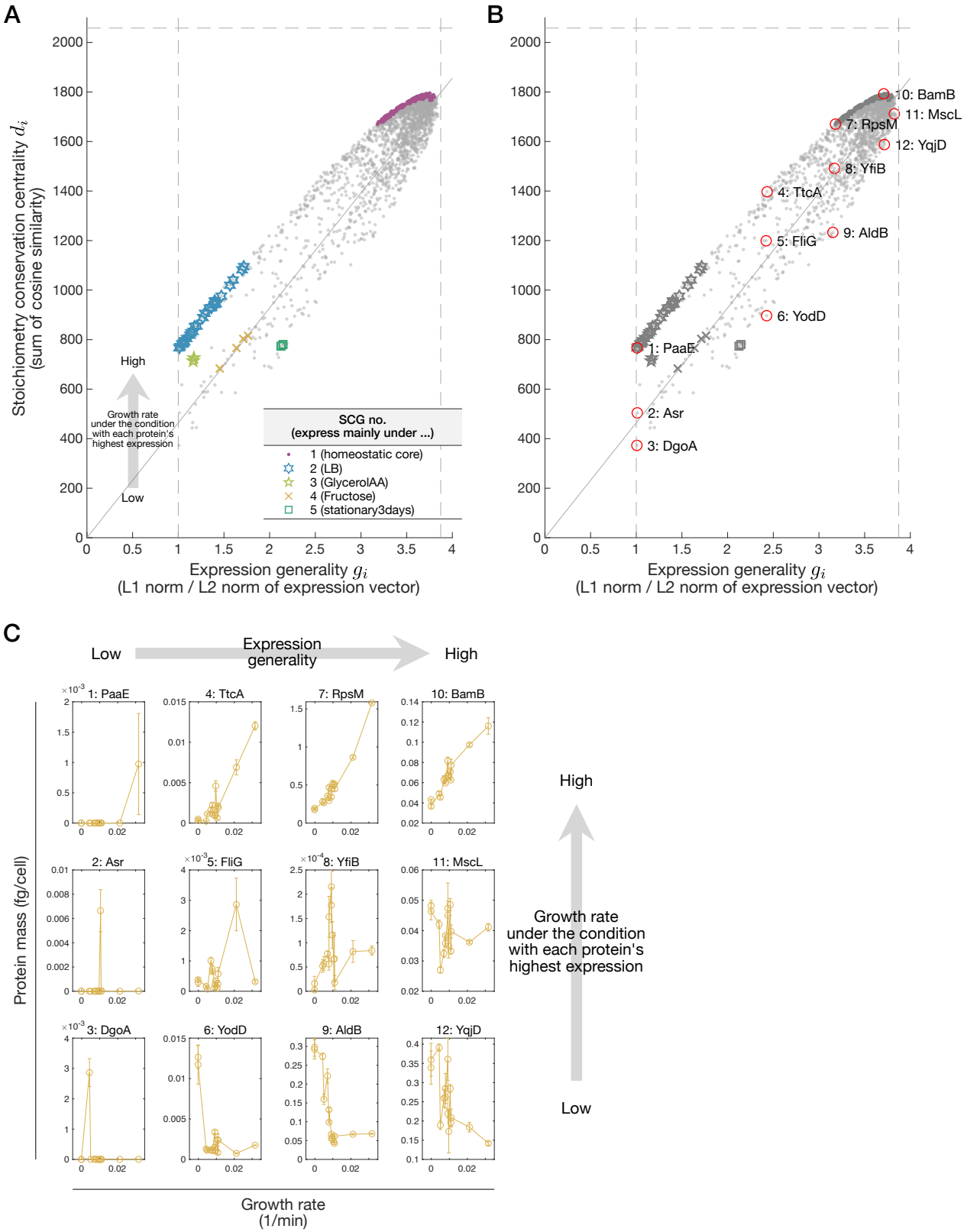
$\Theta = I$ (identity transformation)

→ Commonality between LDA axes (distinction of cellular Raman spectra) and csLE axes (stoichiometric characterization of gene expression patterns) (Supplementary figures)

$\boldsymbol{b}_0 \propto \boldsymbol{b}_0^{\mathrm{est}}$

→ Proportionality between d_i (stoichiometry conservation centrality) and g_i (expression generality) (Fig. 7 and supplementary figures)

Fig. 7


A
B
C
Stoichiometry conservation centrality d_i
(sum of cosine similarity)
Expression generality g_i
(L1 norm / L2 norm of expression vector)
High
Low
Growth rate under the condition with each protein's highest expression
SCG no.
(express mainly under ...)
1 (homeostatic core)
2 (LB)
3 (GlycerolAA)
4 (Fructose)
5 (stationary3days)
1: PaaE
2: Asr
3: DgoA
4: TtcA
5: FliG
6: YodD
7: RpsM
8: YfiB
9: AldB
10: BamB
11: MscL
12: YqjD
Expression generality
Protein mass (fg/cell)
Growth rate
(1/min)

Supplementary Materials

Ken-ichiro F. Kamei,[1*] Koseki J. Kobayashi-Kirschvink,[2]
Takashi Nozoe,[1,3,4] Hidenori Nakaoka,[5] Miki Umetani,[6]
Yuichi Wakamoto[1,3,4*]

[1]Department of Basic Science, Graduate School of Arts and Sciences,
The University of Tokyo, Tokyo 153-8902, Japan
[2]Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
[3]Research Center for Complex Systems Biology, The University of Tokyo,
Tokyo 153-8902, Japan
[4]Universal Biology Institute, The University of Tokyo, Tokyo 153-8902, Japan
[5]Department of Optical Imaging, Advanced Research Promotion Center
Tokushima University, Tokushima 770-8503, Japan
[6]Department of Biology, New York University, New York, NY 10003, USA

*To whom correspondence should be addressed; E-mail:
kenichiro_kamei@cell.c.u-tokyo.ac.jp (K.F.K.), cwaka@mail.ecc.u-tokyo.ac.jp (Y.W.).

Contents

1 Materials and Methods **5**

1.1 Experimental methods, data acquisition, and data analyses 5

1.1.1 Absolute quantitative proteome data 5

1.1.2 *E. coli* strains and culture conditions 6

1.1.3 Raman measurements and preprocessing of spectra 10

1.1.4 Data analysis 11

1.2 Raman-proteome statistical correspondence 12

1.2.1 Notation 12

1.2.2 OLS in LOOCV scheme 12

1.2.3 Permutation test 13

1.3 Characterizing a stoichiometrically-conserved group by analyzing the Raman-proteome correspondence matrix 16

1.3.1 Notation 16

1.3.2 Stoichiometry conservation of ISP COG class 16

1.3.3 On the evaluation of stoichiometry conservation by Pearson's correlation coefficient 18

1.4 Direct characterization of stoichiometrically-conserved groups in omics data 19

1.4.1 Notation 19

1.4.2 Identifying SCGs in omics data 19

1.5 Global proteome structures based on stoichiometric balance 31

1.6 Relevance of centrality of cosine similarity LE structure to biological functions 34

1.6.1 Centrality-essentiality correlation 34

1.6.2 Centrality-evolutionary conservation correlation 36
1.6.3 Centrality-coding/non-coding correlation 37
1.7 Global omics structures characterized by Raman-omics correspondences . . 39
1.7.1 Notation 39
1.7.2 Raman-proteome correspondence matrix as a low-dimensional representation of proteome changes 39
1.8 Evaluating similarity between orthogonal matrix Θ and identity matrix . . 41
1.9 Interpretation of L^1 norm/L^2 norm ratio of an expression vector as a quantitative measure of expression generality 43
1.10 Proteome structure obtained with PCA 46

2 Mathematical analysis and details 47

2.1 Mathematics behind the correspondence 47
2.1.1 Notations 47
2.1.2 Preparations 47
2.1.3 Linear transformation between LDA Raman and proteome 55
2.1.4 Relation between LDA Raman and Ω_B 59
2.1.5 Relation between Ω_B and Ω_{LE} 65
2.2 Quantitative constraint on omics profiles 80
2.2.1 From agreement between $\sqrt{m}\boldsymbol{b}_0$ and $\boldsymbol{b}_0^{\mathrm{est}}$ to proportionality between L^1 norm/L^2 norm ratio and degree 80
2.2.2 Mathematics behind proportionality between stoichiometry conservation centrality d_j and expression generality score g_j 81

3 Extended data analysis 86

3.1 Growth laws 86

3.1.1 Single-gene-level growth law 86
3.1.2 Extended verification of stoichiometry conservation 87
3.1.3 Linear dependence of common abundance ratio on growth rate . . . 88
3.2 Generality of the results 89
3.2.1 Additional datasets with Raman data 89
3.2.2 Additional datasets without Raman data 91

4 Supplementary figures **92**

1 Materials and Methods

1.1 Experimental methods, data acquisition, and data analyses

1.1.1 Absolute quantitative proteome data

We utilized high-quality absolute quantitative proteome data reported by Schmidt *et al.* [1]. In these data, expression levels of more than 55 % of genes of *E. coli* BW25113 strain (more than 95 % of total proteome mass) were quantified under various environmental conditions.

We also used additional absolute quantitative proteome data [1, 2, 3] for checking the generality of our findings (see 3.2). In addition to the proteome data across environmental conditions, Schmidt *et al.* also reported proteomes of *E. coli* strains with different genotype backgrounds (BW25113, MG1655, and NCM3722) cultured in a rich medium or a minimal medium supplemented with glucose. Schubert *et al.* [2] quantified proteomes of *M. tuberculosis* H37Rv strain and *M. bovis* BCG strain under time-course environmental change conditions starting from exponential growth conditions, followed by dormant states induced by decreasing oxygen levels, and finally regrowth conditions with re-aeration. Lahtvee *et al.* [3] quantified proteomes of *S. cerevisiae* under a reference condition and three stressed conditions (ethanol, osmotic pressure, and high temperature, with three stress intensity steps for each type of stress) using chemostat.

For checking the generality of our findings across omics classes, we also used the transcriptome data reported by our previous paper [4]. The data include the transcriptomes of *S. pombe* in rich and minimal media, in nutrient-depleted media, and under various stress conditions.

1.1.2 *E. coli* strains and culture conditions

To quantitatively analyze a linkage between the absolute proteome data generated by Schmidt *et al.* [1] and Raman data, we reproduced the culture conditions used in Schmidt *et al.* [1] as closely as possible in our lab. We obtained three biological replicates.

E. coli strains

We used BW25113, MG1655, and NCM3722 as in Schmidt *et al.* [1]. In particular, BW25113 [5] was used for the main data in this study. The genotype of BW25113 is F^- Δ(*araD-araB*)*567* Δ*lacZ4787*(::rrnB-3) λ^- *rph-1* Δ(*rhaD-rhaB*)*568 hsdR514*, that of MG1655 is F^- λ^- *rph-1*, and that of NCM3722 is F^+, respectively [6, 7, 8].

Culture conditions

We prepared 15 batch culture conditions listed in Table S1. We excluded three culture conditions among the 18 conditions reported in [1] because we could not obtain sufficiently strong cellular Raman signals under those excluded conditions. See [1] for the detail of medium compositions. For "GlucosepH6" medium, 37 % HCl was titrated to the "Glucose" medium. Medium for "stationary1day" and "stationary3days" was the same as "Glucose" medium. LB agar plates were prepared by adding 15 g/L agar to "LB" medium.

Table S1: **List of culture conditions.** (excel)

Cultivation

Culturing *E. coli* cells proceeded in four steps:

Step 1 Growth on LB agar plates.

Cells were taken from a −80 °C glycerol stock and streaked on LB agar plates. The plates were incubated at 37 °C overnight and stored at 4 °C. All subsequent experiments were conducted using colonies on the LB agar plates. Picking colonies from the plates for cultivation was done within four days of storage at 4 °C.

Step 2 Liquid culture under "Glucose" condition.

Several colonies picked from LB agar plates were inoculated into "Glucose" liquid culture medium and grown for about 16 hours. Cells for "Glucose42C" condition were cultured at 42 °C and those for the other conditions were grown at 37 °C.

Step 3 Liquid culture under each condition.

Cells from Step 2 were passaged into each type of medium and grown to exponential phase. Cells for "Glucose42C" condition were grown at 42 °C and those for the other conditions were cultured at 37 °C.

Step 4 Liquid culture under each condition.

Cells from Step 3 were passaged into the respective fresh medium and grown to almost the same level of turbidity as that at the end of Step 3. Cells for "Glucose42C" condition were cultured at 42 °C and those for the other conditions were grown at 37 °C.

For the exponential conditions, cell cultivation was conducted as described above. For the stationary conditions, cultivation of cells at Step 3 was continued instead of proceeding

to Step 4 and ended one or three days after they reached the stationary phase.

The medium volume was 2 mL for all the liquid cultures in our experiments. Borosilicate glass test tubes with a diameter of 16.5 mm and a length of 165 mm were used. A fresh medium was pre-warmed before passage so that its temperature was the same as that of cultivation. All the liquid cultures were under reciprocal shaking at 200 r/min and at an inclination of 45°. Liquid cultures were diluted to an OD_{600} of around 0.01 for passage.

Main differences between our cultivation conditions and those of Schmidt *et al.* [1] are the periods of storage at 4 °C at Step 1 (a maximum of three weeks in [1]), the number of colonies inoculated from plates to liquid medium at the second step (one colony per inoculation in [1]), and medium volumes and shaking conditions of liquid cultures (50 mL liquid culture in 500 mL unbaffled wide-neck Erlenmeyer flasks under orbital shaking at 300 r/min in [1]).

Growth rate measurements

Growth curves were obtained by continuing the Step 3 in cultivation. Cultivation of cells for growth measurements was conducted with 5 mL culture media, not 2 mL, due to a requirement of the device used for continuous turbidity recording (ODBox-C, TAITEC CORPORATION). In addition, cells were washed with each type of fresh medium before inoculation at the beginning of Step 3, and cultivation for growth recording started from an OD_{600} of around 0.001. Growth rates were calculated from the growth curves using the fitting algorithm based on Gaussian processes [9].

1.1.3 Raman measurements and preprocessing of spectra

Cells were washed three times with 0.9 % aqueous solution of NaCl, and 5 µL of the suspension was placed on a synthetic quartz slide glass (TOSHIN RIKO CO., LTD.) and dried. Raman spectra of cells were measured with a Raman microscope (Figure S2), where a custom-built Raman system (STR-Raman, AIRIX) was integrated into a microscope (Ti-E, Nikon). Excitation light was generated by a 532 nm continuous-wave diode-pumped solid-state (DPSS) laser (Gem 532, Laser Quantum). We altered the first version of this Raman microscope [4], and light from the laser oscillator was transmitted by mirrors in this research. A 100× and NA = 0.9 air objective lens (MPLN100X, Olympus) was used. Raman scattered light was collected by an optical fiber and transmitted to a spectrometer (Acton SP2300i, Princeton Instruments). Dispersed light by a 300 gr/mm grating was projected onto an image sensor of an sCMOS camera (OrcaFlash 4.0 v2, Hamamatsu Photonics). The sCMOS camera was water-cooled at 15 °C to reduce dark noise. The exposure time for each cell was 10 s. Randomly selected 15 cells were measured per condition per replicate. Raman spectrum of background was measured for each cell with 10 s exposure in an area close to a targeted cell where neither cells nor NaCl crystals existed.

In our setup, the laser power at the sample stage was 21 mW. The measurement system and processes were controlled by using Micro-Manager 1.4 [10] and a plugin we made.

Readout noise of sCMOS image sensors is pixel-dependent. A noise reduction filter developed in [4] on the basis of [11] was applied to measured spectral images by using 10000 blank images obtained with the same sCMOS sensor with exposure time of 10 s. See [4] for detail.

After noise reduction with the filter, pixel counts were summed up along the direction perpendicular to wavenumber. A background spectrum was subtracted from a cellular Raman spectrum. A pixel region corresponding to the range from $632\,\mathrm{cm}^{-1}$ to $1862\,\mathrm{cm}^{-1}$ was cropped. The cropped spectrum was smoothed with Savitzky–Golay filter [12]. To minimize the effect of laser excitation variations, each spectrum was normalized by subtracting the average and dividing it by the standard deviation.

1.1.4 Data analysis

We wrote scripts and analyzed data using Matlab (R2019a and R2023b), except for Brunner–Munzel test, for which we used R (version 4.0.3) (See 1.6.2).

Related to Figure 2 in the main text, we first performed linear discriminant analysis (LDA) against the Raman data. LDA is a linear classifier; it finds the most discriminatory bases by maximizing the ratio of the between-class variance to the within-class variance, and reduces the dimensions of the data to $m - 1$, where m is the number of classes [13][14][15]. In the case of our main data, classes are culture conditions. In the verification step of the correspondence between the LDA Raman and omics data, we conducted leave-one-out cross-validation (LOOCV). In LOOCV, one condition is used as test data and the remaining conditions are used as training data. This is repeated by changing the condition to exclude.

The detail of the data analyses will follow in the sections below.

1.2 Raman-proteome statistical correspondence

1.2.1 Notation

We write the population-averaged 14-dimensional LDA Raman spectrum vector of each condition as a row vector $\hat{\boldsymbol{r}}_i$ $(i = 1, \ldots, 15)$ and the 2058-dimensional absolute proteome vector of each condition as a row vector $\hat{\boldsymbol{p}}_i$ $(i = 1, \ldots, 15)$. Note that we regarded $\hat{\boldsymbol{r}}_i$ and $\hat{\boldsymbol{p}}_i$ as column vectors in the main text for simple expression of equations.

Our hypothesis of Raman-proteome linear correspondence (Eq. 1 in the main text) is expressed as

$$\hat{\boldsymbol{p}}_i^\top = B \cdot \begin{bmatrix} 1 \\ \hat{\boldsymbol{r}}_i^\top \end{bmatrix}, \tag{1.1}$$

where B is a 2058×15 matrix and $\top$ denotes transpose. In LOOCV, one condition is excluded (let i be the excluded conditoin) and the remaining 14 conditions are used to estimate B. We write the estimated B as B_{-i}^{est}, which is also a 2058×15 matrix. Let $\hat{\boldsymbol{p}}_i^{\mathrm{est}}$ be the estimated proteome of the excluded condition in LOOCV (Figure 2G).

1.2.2 OLS in LOOCV scheme

In the case of LOOCV, 14 $(= 15 - 1)$ conditions are included in a training data. Thus, if all the 14 LDA axes of the low-dimensional Raman data are considered, OLS becomes underdetermined. We excluded higher dimensions of the Raman space to conduct OLS in LOOCV unless otherwise noted. The results described in the main text were obtained by using the first four axes (LDA1 to LDA4). In this case, B_{-i}^{est} is a 2058×5 matrix.

1.2.3 Permutation test

Let a permutation of all the 15 conditions be σ. In our permutation test, we calculated overall estimation errors as $\sum_i \mathrm{dist}\left(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}^{\mathrm{est}}_{\sigma(i)}\right)$, where $\mathrm{dist}\left(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}^{\mathrm{est}}_{\sigma(i)}\right)$ is one of the distance measures between $\hat{\boldsymbol{p}}_i$ and $\hat{\boldsymbol{p}}^{\mathrm{est}}_{\sigma(i)}$ listed in Table S2. There exist 15! sets of σ, and calculating all of them is computationally intensive. Thus, we randomly generated 10^5 permutation sets.

The result presented in the main text is the case where Euclidean metric (PRESS) was used as a distance measure. Likewise, we also obtained small p-values with the other metrics (Table S2).

Table S2: **Evaluation of the overall estimation error with various distance measures (The case where LDA1 to LDA4 axes were used).**
The sum of estimation errors $\sum_i \mathrm{dist}\left(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}_i^{\mathrm{est}}\right)$ was calculated and a permutation test (10^5 permutations) was conducted. In this table, LDA1 to LDA4 axes were used. $\bar{\boldsymbol{x}}$ represents a vector whose all elements are the mean of all elements of $\boldsymbol{x}$. $\boldsymbol{x}_j$ is the j-th element of $\boldsymbol{x}$. $\mathrm{median}_j\ x_j$ represents the median of scalers x_j.

Metric	**Definition of dist$(\boldsymbol{x}, \boldsymbol{y})$**	$\sum_i \mathbf{dist}\left(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}_i^{\mathbf{est}}\right)$	***p*-value**
Square of L^2 norm (PRESS)	$\lVert\boldsymbol{x}-\boldsymbol{y}\rVert_2^2 = \sum_j (\boldsymbol{x}-\boldsymbol{y})_j^2$	2.34×10^3	0.00005
L^1 norm	$\lVert\boldsymbol{x}-\boldsymbol{y}\rVert_1 = \sum_j \left\lvert(\boldsymbol{x}-\boldsymbol{y})_j\right\rvert$	1.40×10^3	0.00002
Cosine distance	$1 - \dfrac{\boldsymbol{x}\cdot\boldsymbol{y}}{\lVert\boldsymbol{x}\rVert_2\lVert\boldsymbol{y}\rVert_2}$	1.52	0.0014
$1-$ Pearson's correlation coefficient	$1 - \dfrac{(\boldsymbol{x}-\bar{\boldsymbol{x}})\cdot(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\lVert\boldsymbol{x}-\bar{\boldsymbol{x}}\rVert_2\lVert\boldsymbol{y}-\bar{\boldsymbol{y}}\rVert_2}$	1.57	0.0012
Median of relative error	$\mathrm{median}_j \dfrac{\left\lvert(\boldsymbol{x}-\boldsymbol{y})_j\right\rvert}{\boldsymbol{x}_j+1}$	0.0536	0.00022

We could also estimate the proteomes with high accuracy using all the 14 dimensions of the LDA space (Table S3). As noted in 1.2.2, the regression is underdetermined in this case. Thus, we simply adopted the minimum-norm solution from among all least-squares solutions.

Table S3: **Evaluation of the overall estimation error with various distance measures (The case where all the 14 LDA axes were used).**
The results obtained by using all the 14 LDA axes are presented. See Table S2 for notations. Note that the system is underdetermined in this case, thus we adopted the minimum-norm solution from among all least-squares solutions.

Metric	**Definition of dist$(\boldsymbol{x}, \boldsymbol{y})$**	$\sum_i \mathbf{dist}\left(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}_i^{\mathbf{est}}\right)$	***p*-value**
Square of L^2 norm (PRESS)	$\lVert\boldsymbol{x}-\boldsymbol{y}\rVert_2^2 = \sum_j (\boldsymbol{x}-\boldsymbol{y})_j^2$	1.63×10^3	0.0019
L^1 norm	$\lVert\boldsymbol{x}-\boldsymbol{y}\rVert_1 = \sum_j \left\lvert(\boldsymbol{x}-\boldsymbol{y})_j\right\rvert$	1.19×10^3	0.00066
Cosine distance	$1 - \dfrac{\boldsymbol{x}\cdot\boldsymbol{y}}{\lVert\boldsymbol{x}\rVert_2\lVert\boldsymbol{y}\rVert_2}$	1.18	0.0879
$1-$ Pearson's correlation coefficient	$1 - \dfrac{(\boldsymbol{x}-\bar{\boldsymbol{x}})\cdot(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\lVert\boldsymbol{x}-\bar{\boldsymbol{x}}\rVert_2\lVert\boldsymbol{y}-\bar{\boldsymbol{y}}\rVert_2}$	1.23	0.085
Median of relative error	$\mathrm{median}_j \dfrac{\left\lvert(\boldsymbol{x}-\boldsymbol{y})_j\right\rvert}{\boldsymbol{x}_j+1}$	0.0418	0.00082

1.3 Characterizing a stoichiometrically-conserved group by analyzing the Raman-proteome correspondence matrix

1.3.1 Notation

The component representation of (1.1) is

$$\overbrace{\begin{pmatrix} p_{i1} \\ p_{i2} \\ \vdots \\ p_{in} \end{pmatrix}}^{\hat{\boldsymbol{p}}_i^\top} = \overbrace{\begin{pmatrix} b_{10} & b_{11} & \cdots & b_{1(m-1)} \\ b_{20} & b_{21} & \cdots & b_{2(m-1)} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n0} & b_{n1} & \cdots & b_{n(m-1)} \end{pmatrix}}^{B} \overbrace{\begin{pmatrix} 1 \\ r_{i1} \\ \vdots \\ r_{i(m-1)} \end{pmatrix}}^{\begin{bmatrix} 1 & \hat{\boldsymbol{r}}_i \end{bmatrix}^\top}, \tag{1.2}$$

where n is the number of proteins, and m is the number of culture conditions. $n = 2058$ and $m = 15$ in our case. Let $\boldsymbol{b}_{h-1}$ be the h-th column of B. For example, $\boldsymbol{b}_0 = \begin{pmatrix} b_{10} & \cdots & b_{n0} \end{pmatrix}^\top$ denotes the constant term for each protein, and $\boldsymbol{b}_1 = \begin{pmatrix} b_{11} & \cdots & b_{n1} \end{pmatrix}^\top$ the coefficient of LDA1 for each protein. The expression level of protein j in the condition i is

$$p_{ij} = b_{j0} + b_{j1} r_{i1} + \ldots + b_{j(m-1)} r_{i(m-1)}. \tag{1.3}$$

1.3.2 Stoichiometry conservation of ISP COG class

In the main text, we revealed that many proteins belonging to "information storage and processing" COG class were aligned on a straight line passing through the origin when the relations between the columns of B were shown in scatter plots (Figure 3A). Consider hypothetical proteins that align perfectly on a straight line through the origin. Let $e_1, \ldots, e_k$ be the indices of such perfectly aligning protein species. Extracting only

these rows for the proteins from (1.2), we obtain

$$\begin{pmatrix} p_{ie_1} \\ p_{ie_2} \\ \vdots \\ p_{ie_k} \end{pmatrix} = \begin{bmatrix} \tilde{\boldsymbol{b}}_0 & \tilde{\boldsymbol{b}}_1 & \cdots & \tilde{\boldsymbol{b}}_{m-1} \end{bmatrix} \begin{pmatrix} 1 \\ r_{i1} \\ \vdots \\ r_{i(m-1)} \end{pmatrix} \tag{1.4}$$

$$= \begin{bmatrix} \tilde{\boldsymbol{b}}_0 & c_1\tilde{\boldsymbol{b}}_0 & \cdots & c_{m-1}\tilde{\boldsymbol{b}}_0 \end{bmatrix} \begin{pmatrix} 1 \\ r_{i1} \\ \vdots \\ r_{i(m-1)} \end{pmatrix} \tag{1.5}$$

$$= \left(1 + c_1 r_{i1} + \cdots + c_{m-1} r_{i(m-1)}\right)\tilde{\boldsymbol{b}}_0, \tag{1.6}$$

where c_i $(i = 1, 2, \ldots, m-1)$ are constants and $\tilde{\boldsymbol{b}}_h \coloneqq \begin{pmatrix} b_{e_1 h} & \cdots & b_{e_k h} \end{pmatrix}^\top$. For our data, Figures S4A and S4B correspond to (1.6). In these plots, the y-axis represents $\hat{\boldsymbol{p}}_i^\top$, and the x-axis $\tilde{\boldsymbol{b}}_0$. Many "Information storage and processing" proteins indeed align on a straight line through the origin with different slopes for different conditions (Figure S4A). In contrast, many proteins in other COG classes do not align on a straight line (Figure S4B).

Importantly, for a pair of proteins e_α, e_β that align on the straight line,

$$\frac{p_{ie_\alpha}}{p_{ie_\beta}} = \frac{b_{e_\alpha 0}}{b_{e_\beta 0}} \tag{1.7}$$

holds from (1.6). The right-hand side of (1.7) does not contain condition index i, which means that the abundance ratio of the proteins remains constant independently of the conditions.

1.3.3 On the evaluation of stoichiometry conservation by Pearson's correlation coefficient

In the main text, we used Pearson's correlation coefficients to confirm the stoichiometry conservation of many ISP COG class members (Figure 3B). However, strictly speaking, cosine similarity is a more appropriate measure to evaluate stoichiometry conservation. In this analysis, cosine similarity can be written as

$$\cos\theta_{\tilde{\hat{\boldsymbol{p}}}_i \tilde{\hat{\boldsymbol{p}}}_j} = \frac{\tilde{\hat{\boldsymbol{p}}}_i \cdot \tilde{\hat{\boldsymbol{p}}}_j}{\left\|\tilde{\hat{\boldsymbol{p}}}_i\right\|_2 \left\|\tilde{\hat{\boldsymbol{p}}}_j\right\|_2}, \tag{1.8}$$

where $\tilde{\hat{\boldsymbol{p}}}_i$ and $\tilde{\hat{\boldsymbol{p}}}_j$ are the vectors representing the protein abundance for the proteome subgroups ("Information storage and processing" COG class, "Cellular processes and signaling" COG class, and "Metabolism" COG class) for conditions i and j, respectively ($1 \leq i, j \leq m$). Cosine similarity version of Figure 3B is Figure S4F. The cosine similarity takes the maximum value 1 only when abundance ratios between all considered proteins are perfectly the same between compared two conditions.

In addition, we also examined differences between COG classes by calculating Pearson's correlation coefficients of log abundances (Figure S4E).

1.4 Direct characterization of stoichiometrically-conserved groups in omics data

1.4.1 Notation

Let $\boldsymbol{p}_i$ be a column vector representing the abundances of protein i. Each component of this vector indicates the abundance of protein i under each condition. Therefore,

$$\boldsymbol{p}_i = \begin{pmatrix} p_{1i} \\ p_{2i} \\ \vdots \\ p_{mi} \end{pmatrix}, \tag{1.9}$$

where $m = 15$ in our case. Note that $\boldsymbol{p}_i$ defined here is a 15-dimensional column vector and different from $\hat{\boldsymbol{p}}_i$ introduced previously, which was a 2058-dimensional row vector.

1.4.2 Identifying SCGs in omics data

As explained in the main text, we extracted stoichiometrically-conserved groups (SCGs) directly from the omics data without referring to Raman data nor COG classification. We evaluated similarity of expression patterns for all the combinations of proteins using cosine similarity. Specifically, cosine similarity between proteins i and j is calculated as

$$\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} := \frac{\boldsymbol{p}_i \cdot \boldsymbol{p}_j}{\|\boldsymbol{p}_i\|_2 \|\boldsymbol{p}_j\|_2}. \tag{1.10}$$

This is the inner product of normalized $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. Note that $0 \leq \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} \leq 1$ as protein abundances of any proteins are non-negative. $\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}$ takes the maximum value 1 if and only if $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ point in identical direction, i.e., the abundance ratios of proteins

i and j are constant under all the conditions. Therefore, if we extract only the proteins connected with high cosine similarity from all $\binom{n}{2}$ protein pairs, they would constitute proteome fractions in each of which the abundance ratios of the proteins remain almost constant across all the m conditions. We hence extracted only the protein pairs whose cosine similarity was above a high threshold of 0.995. As a result, we obtained several SCGs, in each of which the protein species are linked to each other with high cosine similarity (Figure 4B and 4C).

The genes in each SCG are listed in Tables S4, S5, S7, S8 and S6. Note that there are many other minor components (Figure 4B), some of which may have an expression pattern similar to another component but are separated due to the high threshold.

The positions of members of the SCGs on the chromosome are shown in Figure 4D (SCG 1(homeostatic core)) and Figure S5E (SCGs 2–5).

Table S4: **Gene list of SCG 1 (homeostatic core).**

Members of homeostatic core (Figure 4, cosine similarity threshold: 0.995). The description of each gene is cited from [1].

Name	Description
rpoC	DNA-directed RNA polymerase subunit beta'
rpoB	DNA-directed RNA polymerase subunit beta
tufA	Elongation factor Tu 1
infB	Translation initiation factor IF-2
fusA	Elongation factor G
glyS	Glycyl-tRNA synthetase beta subunit
rpsA	30S ribosomal protein S1
leuS	Leucyl-tRNA synthetase
pheT	Phenylalanyl-tRNA synthetase beta chain
aspS	Aspartyl-tRNA synthetase
valS	Valyl-tRNA synthetase
secA	Protein translocase subunit secA
gyrA	DNA gyrase subunit A
pepN	Aminopeptidase N
tsf	Elongation factor Ts
tig	Trigger factor
pta	Phosphate acetyltransferase
bamA	Outer membrane protein assembly factor yaeT
rne	Ribonuclease E
ftsZ	Cell division protein ftsZ
gyrB	DNA gyrase subunit B
polA	DNA polymerase I
rplB	50S ribosomal protein L2
prlC	Oligopeptidase A

Gene list of SCG 1 continued on next page

Name	Description
rho	Transcription termination factor Rho
ftsH	ATP-dependent zinc metalloprotease FtsH
nusA	Transcription elongation protein nusA
lysS	Lysyl-tRNA synthetase
metG	Methionyl-tRNA synthetase
glnS	Glutaminyl-tRNA synthetase
lpdA	Dihydrolipoyl dehydrogenase
serS	Seryl-tRNA synthetase
surA	Chaperone surA
rpsB	30S ribosomal protein S2
gltX	Glutamyl-tRNA synthetase
lptD	LPS-assembly protein lptD
argS	Arginyl-tRNA synthetase
fabB	3-oxoacyl-[acyl-carrier-protein] synthase 1
pheS	Phenylalanyl-tRNA synthetase alpha chain
clpX	ATP-dependent Clp protease ATP-binding subunit ClpX
accC	Biotin carboxylase
pyrG	CTP synthase
tolC	Outer membrane protein tolC
rplE	50S ribosomal protein L5
accA	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha
hflK	Modulator of FtsH protease HflK
pdxB	Erythronate-4-phosphate dehydrogenase
ygfZ	tRNA-modifying protein ygfZ
pmbA	Protein pmbA
rplA	50S ribosomal protein L1
hldD	ADP-L-glycero-D-manno-heptose-6-epimerase
mreB	Rod shape-determining protein mreB

Gene list of SCG 1 continued on next page

Name	Description
acrA	Acriflavine resistance protein A
gor	Glutathione reductase
hisS	Histidyl-tRNA synthetase
rpsC	30S ribosomal protein S3
glmM	Phosphoglucosamine mutase
lepA	Elongation factor 4
ffh	Signal recognition particle protein
secD	Protein-export membrane protein secD
lpoA	Penicillin-binding protein activator LpoA
rhlB	ATP-dependent RNA helicase rhlB
rpsG	30S ribosomal protein S7
rpsD	30S ribosomal protein S4
minD	Septum site-determining protein minD
cyoA	Ubiquinol oxidase subunit 2
mdoG	Glucans biosynthesis protein G
rplC	50S ribosomal protein L3
glmU	Bifunctional protein glmU
rpsF	30S ribosomal protein S6
rpsE	30S ribosomal protein S5
hemL	Glutamate-1-semialdehyde 2,1-aminomutase
hldE	Bifunctional protein hldE
ubiE	Ubiquinone/menaquinone biosynthesis methyltransferase ubiE
sspA	Stringent starvation protein A
nusG	Transcription antitermination protein nusG
prfB	Peptide chain release factor 2
dacA	D-alanyl-D-alanine carboxypeptidase dacA
rplF	50S ribosomal protein L6
fabG	3-oxoacyl-[acyl-carrier-protein] reductase

Gene list of SCG 1 continued on next page

Name	Description
ftsY	Cell division protein ftsY
dcrB	Protein dcrB
mlaC	Probable phospholipid-binding protein mlaC
hflC	Modulator of FtsH protease HflC
coaB	Coenzyme A biosynthesis bifunctional protein CoaBC
ybiT	Uncharacterized ABC transporter ATP-binding protein YbiT
oxyR	Hydrogen peroxide-inducible genes activator
rpsH	30S ribosomal protein S8
fkpA	FKBP-type peptidyl-prolyl cis-trans isomerase fkpA
frr	Ribosome-recycling factor
fabD	Malonyl CoA-acyl carrier protein transacylase
hslO	33 kDa chaperonin
ybeZ	PhoH-like protein
hemX	Putative uroporphyrinogen-III C-methyltransferase
rplY	50S ribosomal protein L25
rplK	50S ribosomal protein L11
rpsI	30S ribosomal protein S9
bamB	Lipoprotein yfgL
bamD	UPF0169 lipoprotein yfiO
kdgR	Transcriptional regulator kdgR
glnD	[Protein-PII] uridylyltransferase
yniC	Phosphatase YniC
rpsJ	30S ribosomal protein S10
rplX	50S ribosomal protein L24
rplD	50S ribosomal protein L4
rplQ	50S ribosomal protein L17
ppa	Inorganic pyrophosphatase
rpsM	30S ribosomal protein S13

Gene list of SCG 1 continued on next page

Name	Description
rplN	50S ribosomal protein L14
ybaB	UPF0133 protein ybaB
yidC	Inner membrane protein oxaA
lptB	Lipopolysaccharide export system ATP-binding protein LptB
suhB	Inositol-1-monophosphatase
yejK	Nucleoid-associated protein YejK
ghrA	Glyoxylate/hydroxypyruvate reductase A
rsmI	Ribosomal RNA small subunit methyltransferase I
hemY	Protein hemY
uup	ABC transporter ATP-binding protein uup
hrpA	ATP-dependent RNA helicase hrpA
rplJ	50S ribosomal protein L10
rplM	50S ribosomal protein L13
fur	Ferric uptake regulation protein
rplS	50S ribosomal protein L19
rcsB	Capsular synthesis regulator component B
mrp	Protein mrp
glyQ	Glycyl-tRNA synthetase alpha subunit
greA	Transcription elongation factor greA
nrdB	Ribonucleoside-diphosphate reductase 1 subunit beta
wbbI	Uncharacterized protein yefG
udk	Uridine kinase
mnmG	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG
rplL	50S ribosomal protein L7/L12
rplI	50S ribosomal protein L9
rpoZ	DNA-directed RNA polymerase subunit omega
ybbN	Uncharacterized protein ybbN
yfiF	Uncharacterized tRNA/rRNA methyltransferase YfiF

Gene list of SCG 1 continued on next page

Name	Description
yedD	Uncharacterized lipoprotein yedD
rpmD	50S ribosomal protein L30
tatB	Sec-independent protein translocase protein tatB
yfgM	UPF0070 protein yfgM
kdsB	3-deoxy-manno-octulosonate cytidylyltransferase
rpoN	RNA polymerase sigma-54 factor
fdx	2Fe-2S ferredoxin
rplV	50S ribosomal protein L22
rplO	50S ribosomal protein L15
fabZ	(3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase
mipA	MltA-interacting protein
ssb	Single-stranded DNA-binding protein
yiaF	Uncharacterized protein yiaF
secY	Preprotein translocase subunit SecY
rbfA	Ribosome-binding factor A
potA	Spermidine/putrescine import ATP-binding protein PotA
rimM	Ribosome maturation factor rimM
trxA	Thioredoxin-1
rpsS	30S ribosomal protein S19
rpsU	30S ribosomal protein S21
accB	Biotin carboxyl carrier protein of acetyl-CoA carboxylase
engB	Probable GTP-binding protein engB
tatA	Sec-independent protein translocase protein tatA
rfbD	dTDP-4-dehydrorhamnose reductase
ribF	Riboflavin biosynthesis protein ribF
folP	Dihydropteroate synthase
lepB	Signal peptidase I
sspB	Stringent starvation protein B

Gene list of SCG 1 continued on next page

Name	Description
hupA	DNA-binding protein HU-alpha
rpsP	30S ribosomal protein S16
rplP	50S ribosomal protein L16
rpsT	30S ribosomal protein S20
rpsK	30S ribosomal protein S11
rplU	50S ribosomal protein L21
rplR	50S ribosomal protein L18
lpxA	Acyl-[acyl-carrier-protein]–UDP-N-acetylglucosamine O-acyltransferase
yceD	Uncharacterized protein yceD
queC	7-cyano-7-deazaguanine synthase
rpmA	50S ribosomal protein L27
rpmG	50S ribosomal protein L33
rpmF	50S ribosomal protein L32
rpsN	30S ribosomal protein S14
rplT	50S ribosomal protein L20
nudK	GDP-mannose pyrophosphatase nudK
rplW	50S ribosomal protein L23
trmB	tRNA (guanine-N(7)-)-methyltransferase
rluB	Ribosomal large subunit pseudouridine synthase B
rpsR	30S ribosomal protein S18
secG	Protein-export membrane protein SecG
rlmE	Ribosomal RNA large subunit methyltransferase E
yfaY	CinA-like protein
trmA	tRNA (uracil-5-)-methyltransferase
rpmH	50S ribosomal protein L34
yajC	UPF0092 membrane protein yajC
yheU	UPF0270 protein yheU

Gene list of SCG 1 end

Table S5: **Gene list of SCG 2.**

Members in SCG 2 (Figure 4, cosine similarity threshold: 0.995).

The description of each gene is cited from [1].

Name	**Description**
fdoG	Formate dehydrogenase-O major subunit
dsdA	D-serine dehydratase
treC	Trehalose-6-phosphate hydrolase
sdaB	L-serine dehydratase 2
nanA	N-acetylneuraminate lyase
garD	D-galactarate dehydratase
proV	Glycine betaine/L-proline transport ATP-binding protein ProV
garR	2-hydroxy-3-oxopropionate reductase
nanK	N-acetylmannosamine kinase
fdoH	Formate dehydrogenase-O iron-sulfur subunit
aphA	Class B acid phosphatase
nanE	Putative N-acetylmannosamine-6-phosphate 2-epimerase
srlB	Glucitol/sorbitol-specific phosphotransferase enzyme IIA component
ibpB	Small heat shock protein ibpB
hybC	Hydrogenase-2 large chain
proW	Glycine betaine/L-proline transport system permease protein proW
srlE	Glucitol/sorbitol-specific phosphotransferase enzyme IIB component
fdoI	Formate dehydrogenase, cytochrome b556(fdo) subunit
preT	Uncharacterized oxidoreductase yeiT
garL	5-keto-4-deoxy-D-glucarate aldolase
paaB	Phenylacetic acid degradation protein paaB
paaK	Phenylacetate-coenzyme A ligase
paaE	Probable phenylacetic acid degradation NADH oxidoreductase paaE
ykgE	Uncharacterized protein ykgE

Gene list of SCG 2 continued on next page

Name	Description
ybjT	Uncharacterized protein ybjT
ykgG	Uncharacterized protein ykgG

Gene list of SCG 2 end

Table S6: **Gene list of SCG 3.**

Members in SCG 3 (Figure 4, cosine similarity threshold: 0.995).

The description of each gene is cited from [1].

Name	Description
wzc	Tyrosine-protein kinase wzc
amiC	N-acetylmuramoyl-L-alanine amidase AmiC

Table S7: **Gene list of SCG 4.**

Members in SCG 4 (Figure 4, cosine similarity threshold: 0.995).

The description of each gene is cited from [1].

Name	Description
fruB	Multiphosphoryl transfer protein
fruK	1-phosphofructokinase
fruA	PTS system fructose-specific EIIBC component
narI	Respiratory nitrate reductase 1 gamma chain

Table S8: **Gene list of SCG 5.**

Members in SCG 5 (Figure 4, cosine similarity threshold: 0.995).

The description of each gene is cited from [1].

Name	**Description**
hdeB	Protein hdeB
hdeA	Chaperone-like protein hdeA

1.5 Global proteome structures based on stoichiometric balance

In the previous section, we identified SCGs by setting a threshold of cosine similarity for extracting protein pairs. We next removed the threshold and considered the "distance" with respect to cosine similarity for all the protein pairs to capture the global proteome structure including SCGs.

The cosine similarity for all the $\binom{n}{2}$ pairs of proteins can be summarized in one matrix as

$$A := \left(\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}\right)_{ij} = \begin{pmatrix} \cos\theta_{\boldsymbol{p}_1\boldsymbol{p}_1} & \cdots & \cos\theta_{\boldsymbol{p}_1\boldsymbol{p}_n} \\ \vdots & \ddots & \vdots \\ \cos\theta_{\boldsymbol{p}_n\boldsymbol{p}_1} & \cdots & \cos\theta_{\boldsymbol{p}_n\boldsymbol{p}_n} \end{pmatrix}, \tag{1.11}$$

where (i, j) component represents cosine similarity between proteins i and j. Assuming that this matrix is an adjacency matrix in graph theory and network theory, the entire proteomes are considered as a weighted undirected complete graph (with loops), where nodes correspond to protein types and any protein pair is connected by an undirected edge. Each edge is weighted by the cosine similarity between the two protein species at both ends. Note that all the diagonal elements of A are one, which represents that the each node has a loop with weight of one. These were introduced just for simplicity.

To ask whether the SCGs identified in the previous section have any unique features in this network, we evaluated the degree to which each node is central in the network structure. In graph theory, "centrality" is known as an index to measure how "important" or "influential" each node is. In particular, we employed a measure called "degree centrality" (for weighted graphs) [16][17]. Degree centrality, which is also called "degree", simply measures "influence" of a node on a network on the basis of links with its direct neighborhood. One can obtain a degree centrality value by calculating the sum of the

weights of all the edges connected to each node (see also the definition of the degree matrix D (1.13) below). We note that in our graph, degree centrality vector $A\mathbf{1}_n (= D\mathbf{1}_n)$, where $\mathbf{1}_n$ is an n-dimensional column vector of which all elements are one, is equal to the eigenvector corresponding to the largest eigenvalue of a "normalized" adjacency matrix $(D^{-1}A)^\top = AD^{-1}$ up to multiplication by a constant. From this perspective, the centrality index we adopted measures "influence" of a node in a recursive manner depending on "influence" of its neighboring nodes. A well-known example of this centrality indicator is Google's PageRank [18] used for ranking web pages on the World Wide Web. It can be regarded as a variant of "eigenvector centrality" (the eigenvector corresponding to the largest eigenvalue of the adjacency matrix A) [19][17]. As explained in the main text, the protein species in the homeostatic core (the largest SCG) had high centrality scores, while those in the other condition-specific SCGs had low centrality scores (Figure 5A).

We directly observed the global stoichiometry conservation structure of this proteome graph using Laplacian eigenmaps (Figures 5K, 5L, and 6D). In general, a graph can be uniquely specified not only by the adjacency matrix A, but also by the Laplacian matrix L defined as

$$L \coloneqq D - A, \tag{1.12}$$

where $D = (d_{ij})$ is the degree matrix with the components of

$$d_{ij} = \begin{cases} (A\mathbf{1}_n)_i & (i = j) \\ 0 & (i \neq j) \end{cases}, \tag{1.13}$$

where $\mathbf{1}_n$ is an n-dimensional column vector of which all elements are one. The (i, i)-element of D represents the sum of the weights of all the edges connected to node i. In our case, it represents the sum of cosine similarity values between protein i and the

other proteins. To see the entire proteome graph structure, we specifically employed the normalized Laplacian,

$$L_{\mathrm{rw}} = D^{-1}L = I - D^{-1}A. \tag{1.14}$$

We remark that there are two types of often-used normalized Laplacian matrices, L_{rw} and $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$, in the field of machine learning [20], and our mathematical analysis can provide a clear interpretation to each of them in the context of the Raman-proteome linear correspondence as described in 2.1.5.

There exist $m-1$ non-trivial eigenvalues of L_{rw} that are greater than zero and less than one. We write these $m-1$ eigenvalues as $\lambda_{\mathrm{LE}1}, \ldots, \lambda_{\mathrm{LE}(m-1)}$ from the smallest and the corresponding eigenvectors as $\boldsymbol{v}_{\mathrm{rw},1}, \ldots, \boldsymbol{v}_{\mathrm{rw},(m-1)}$. Additionally, we denote the eigenvector corresponding to the eigenvalue zero as $\boldsymbol{v}_{\mathrm{rw},0}$. Using these eigenvectors, one can construct a matrix $\tilde{V}_{\mathrm{rw}} = \begin{bmatrix} \boldsymbol{v}_{\mathrm{rw},0} & \boldsymbol{v}_{\mathrm{rw},1} & \cdots & \boldsymbol{v}_{\mathrm{rw},(m-1)} \end{bmatrix}$ and visualize a proteome, assigning protein j with a coordinate specified by the elements after the second column in the j-th row of $\tilde{V}_{\mathrm{rw}}$, i.e. by the j-th row of $\begin{bmatrix} \boldsymbol{v}_{\mathrm{rw},1} & \cdots & \boldsymbol{v}_{\mathrm{rw},(m-1)} \end{bmatrix}$. The cosine similarity structure we illustrate in the main figures was produced by using these eigenvectors. For example, csLE1-csLE2 figure in the main text (Figure 6D) is a scatter plot between $\boldsymbol{v}_{\mathrm{rw},1}$ and $\boldsymbol{v}_{\mathrm{rw},2}$. Note that the closer to one the cosine similarity of a protein pair is (the more similar their expression patterns are), the "closer" the two protein species are placed (See 2.1.5 for details).

This method of obtaining low-dimensional representation of data using eigenvectors of a graph Laplacian is known as Laplacian eigenmaps [21][22]. Thus, what we explained above is the Laplacian eigenmaps of a graph with edges weighted with cosine similarity of expression patterns of nodes (protein species). It differs from the original and common usages of Laplacian eigenmaps in that the graph we considered is a complete graph (with

loops) and that the weight of edges (pairwise similarity of nodes) is cosine similarity. It has made all the mathematical formulations linear, which allowed us to biologically interpret the results with mathematically rigorous analyses. We also remark that our graph representation of proteome does not rely on existing knowledge on the underlying interaction and regulatory networks of proteins and is based only on final expression levels of the proteins. Therefore, the results are robust against the uncertainty of underlying molecular detail.

1.6 Relevance of centrality of cosine similarity LE structure to biological functions

1.6.1 Centrality-essentiality correlation

As mentioned in the main text, centrality of protein species with regard to stoichiometry conservation correlates with gene essentiality (Figure 5B). We analyzed the proteome data from all the 22 conditions reported by [1] in Figure 5B. Interestingly, the centrality-essentiality correlation becomes weaker when the analysis was conducted with the data from fewer conditions (Figure S6A).

We obtained the list of essential genes of *E. coli* from EcoCyc [23] on September 23rd, 2020. The list contained 318 essential genes in total. The essentiality of the genes in this list was determined on the basis of whether single-gene knockouts of BW25113 (Keio collection) could grow under LB condition at 37 °C [6].

We also confirmed centrality-essentiality correlation for *S. pombe* transcriptome data [4] (Figure S6B, see 3.2). For this analysis, we downloaded the list of essential genes of *S. pombe* from PomBase [24] on May 13th, 2022. The list contained 1221 essential genes in total. Here, the essentiality data by PomBase was based on the Fission Yeast

Phenotype Ontology terms "inviable vegetative cell population" (FYPO:0002061) and "viable vegetative cell population" (FYPO:0002060) [25]. Note that in our *S. pombe* essentiality analysis, we focused only on coding genes, whereas the cosine similarity LE structure was calculated using both coding and non-coding genes. See 1.6.3 and Figure S6C for the proportion of coding genes in each bin in Figure S6B. Eleven coding genes in the *S. pombe* transcriptome data were not found in current PomBase. Thus, some bins do not show 100% in total in Figure S6B.

Stoichiometry conservation centrality in human cells was evaluated using two kinds of *H. sapiens* transcriptome data: (i) human cell atlas data reported in [26] (Figure 5F) and (ii) genome-wide Perturb-seq data reported in [27] (Figure 5H).

The human cell atlas data [26] contain gene expression profiles in cells from 15 fetal organs. To calculate stoichiometry conservation centrality from the human cell atlas, we analyzed the pseudobulk data (GSE156793_S4_gene_expression_tissue.txt provided at the Gene Expression Omnibus). We calculated stoichiometry conservation centrality value of each gene using expression level data of 53,908 genes that are expressed at least in one organ.

The Perturb-seq data we used are gene expression profiles in a chronic myeloid leukemia cell line K562 [27]. This dataset contains single-cell RNA sequencing data of genetically perturbed cells in which expression of targeted genes is inhibited by CRISPR interference. We analyzed the pseudobulk data (K562_gwps_raw_bulk_01.h5ad provided at Figshare) to calculate stoichiometry conservation centrality. We evaluated stoichiometry conservation centrality value of each gene using the expression data of all the 8,248 genes in the Perturb-seq data. We remark that this dataset did not contain genes that showed no expression under all the reported genetic perturbation conditions.

Human gene essentiality was determined by referring to another dataset reported in

[28], in which fitness cost imposed by gene inactivation was evaluated by a CRISPR-based method [28]. The fitness cost was quantified by an index called CRISPR score; genes with lower CRISPR scores are considered more essential [28]. We used the CRISPR scores calculated with a human chronic myelogenous leukemia cell line KBM7.

The CRISPR scores of 16,996 genes and 7,462 genes were found in [28] among the genes whose stoichiometry conservation centrality was evaluated using the human cell atlas data [26] and the Perturb-seq data [27], respectively. We evaluated the correlations between stoichiometry conservation centrality and gene essentiality (CRISPR scores) for these common genes in Figure 5F and 5H. The correlations were examined with Brunner–Munzel test [29] using R (version 4.0.3) and "brunnermunzel" package (version 2.0) [30].

1.6.2 Centrality-evolutionary conservation correlation

As mentioned in the main text, centrality of proteins with regard to expression stoichiometry conservation weakly correlates with evolutionary conservation represented by the number of orthologs based on protein sequences (Figure 5C). In Figure 5C, we analyzed the proteome data from all the 22 conditions reported in [1]. We also confirmed the relation for the *E. coli* proteome data from fewer conditions which we had used for our Raman-proteome correspondence analyses (Figure S6D).

We obtained the ortholog data from OrthoMCL-DB [31] (release 6.12). We used the number of orthologs in all of the "Core species" and the "Peripheral species" of OrthoMCL, which are across the three domains (Bacteria, Archaea, and Eukaryota), as a proxy for evolutionary conservation of each protein. To examine the correlation, we performed Brunner–Munzel test [29] using R (version 4.0.3) and "brunnermunzel" package (version 2.0) [30]. The *E. coli* proteome data contain 15 proteins with IDs which were not found in OrthoMCL-DB for technical reasons such as changes in IDs in the past

and thus, we manually processed these 15 proteins.

We also examined *S. pombe* transcriptome data [4] (Figures S6E–G, see 3.2). We obtained ortholog data from OrthoMCL-DB [31] (release 6.12). The *S. pombe* transcriptome data have 11 coding genes which were not found in both current PomBase and OrthoMCL-DB, and two coding genes which were found in PomBase but not in OrthoMCL-DB. The *S. pombe* transcriptome data contain not only coding genes but also non-coding genes, and we obtained the cosine similarity LE structure using both.

We also evaluated stoichiometry conservation-evolutionary conservation correlation using the human cell atlas data [26] (Figure 5G) and the genome-wide Perturb-seq data [27] (Figure 5I). Ortholog data for these analyses were obtained from OrthoMCL-DB (release 6.20). We found the ortholog data in OrthoMCL-DB for 18,959 genes among the 53,908 genes with stoichiometry conservation centrality evaluated with the human cell atlas data. We remark that 98.7% of the 18,959 genes were classified as coding genes in the human cell atlas data. We also found the ortholog data for 7,957 genes among the 8,248 genes with stoichiometry conservation centrality evaluated with the Perturb-seq data. The correlations were examined with Brunner–Munzel test [29] using R (version 4.0.3) and "brunnermunzel" package (version 2.0) [30].

1.6.3 Centrality-coding/non-coding correlation

As mentioned in the main text and 1.6.1, centrality of genes with regard to stoichiometry conservation clearly correlates with coding/non-coding classification of genes in *S. pombe*. We observed this trend using *S. pombe* transcriptome data [4] (Figure S6C). The coding/non-coding assignment of each gene is based on PomBase [24] data downloaded on October 11th, 2022.

We observed a comparable correlation even in the human cell atlas data (Figure 5E).

The gene type assignment is based on the human cell atlas data. Note that almost all the genes in the Perturb-seq data were coding genes.

1.7 Global omics structures characterized by Raman-omics correspondences

1.7.1 Notation

Let $\hat{\boldsymbol{b}}_j$ denote the j-th row in B (see Equtation (1.2)). It is an m-dimensional row vector whose components represent coefficients of protein j. The first component is the constant term, and the i-th component is the coefficient for LDA$(i-1)$ Raman. Below we consider the coefficients normalized with the constant terms,

$$\hat{\boldsymbol{b}}_j^{\text{norm}} := \begin{pmatrix} 1 & \frac{b_{j1}}{b_{j0}} & \dots & \frac{b_{j(m-1)}}{b_{j0}} \end{pmatrix}. \tag{1.15}$$

1.7.2 Raman-proteome correspondence matrix as a low-dimensional representation of proteome changes

We asked whether the stoichiometry conservation structure of the proteomes revealed by LE (Figures 5K, 5L and 6D) is relevant to the low-dimensional Raman LDA space. To address this, we focused on a proteome low-dimensional structure specified by the Raman-proteome coefficients, motivated by the fact that the analysis of B led to the discovery of a proteome fraction that conserves mutual stoichiometry (Figure 3). We considered a space where $\hat{\boldsymbol{b}}_j^{\text{norm}}$ represents the coordinate of each protein. From (1.6) or (1.7), protein species whose abundance ratios remain constant have an identical coordinate in this normalized coefficient space. The proteome in this normalized coefficient space is shown in Figures 6A and 6C.

This structure (Figures 6A and 6C) is constructed by using the Raman LDA axes (dual basis) and is different from the cosine similarity LE structure (Figures 5K, 5L and

6D), which is independent of Raman information. Therefore, it is nontrivial that these two structures are similar. This similarity suggests that differences in cellular Raman spectra captured by LDA might be quantitatively related to the omics structure deduced from stoichiometry-conserving relations. We will mathematically analyze this similarity in the section 2.

1.8 Evaluating similarity between orthogonal matrix Θ and identity matrix

As we see in 2.1.5, an orthogonal matrix Θ that appears in the relation connecting the two types of proteome structure must be close to an identity matrix to guarantee the structural similarity. To evaluate to what extent Θ is close to an identity matrix, we generated many random orthogonal matrices [32] and compared Θ and the identity matrix with them.

We first multiplied each orthogonal matrix by itself in the sense of Hadamard product (element-wise product). Then, we regarded the resultant matrix as a scatter plot (Figure S9B) and calculated its Pearson correlation coefficient, assuming that (i, j) element was the frequency of "data points" at the coordinate (i, j). The obtained Pearson correlation coefficient can be regarded as a measure of closeness to the identity matrix. In the case of identity matrix, the correlation coefficient takes the maximum value, one, because non-zero values are concentrated on the diagonal part.

We calculated the square of each matrix in the sense of Hadamard product for two reasons. First, since all elements of the resultant matrix are non-negative, one can ensure that the number of "points" is non-negative at any coordinate. Note that it is not necessarily an integer here. Second, the sum of all the elements of the resultant matrix is necessarily m. Thus, the total number of "points" is equally m for any $m \times m$ orthogonal matrices compared.

In addition to this method, we also evaluated the closeness of Θ to the identity matrix (i) by comparing the magnitudes of off-diagonal elements among Θ, the identity matrix, and random orthogonal matrices, and (ii) by comparing the magnitudes of elements of leading principal submatrices among Θ, the identity matrix, and random orthogonal matrices. In (i), from a part consisting of $(m-1)$- and $-(m-1)$-diagonals ($(1, m)$ and $(m, 1)$

elements) to the whole matrix, we expand step by step the area to consider by including i- and $-i$-diagonals ($m-1 \geq i \geq 0$, the final step is inclusion of the main diagonal), and calculated the sum of the square of the elements in the area at each step. In (ii), from the smallest leading principal submatrix ($(1,1)$ element) to the whole matrix, we expand the area to consider step by step, and calculated the sum of the square of the elements in the area at each step. See also schematic diagrams in the supplementary figures (e.g. Figures S9D and S9E).

1.9 Interpretation of L^1 norm/L^2 norm ratio of an expression vector as a quantitative measure of expression generality

In 2.1.5, we will also see that even if Θ is close to the identity matrix, there is another condition which must be met to guarantee the similarity of the two types of proteome structure. By considering the mathematics behind the condition, we will reveal that the two indices, stoichiometry conservation centrality (degree) $d_j = \sum_i \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}$ and expression generality score $g_j = \left\|\boldsymbol{p}_j\right\|_1 / \left\|\boldsymbol{p}_j\right\|_2$ must be mutually proportional. Here, we explain why g_j is a quantitative measure of the generality (or constancy) of expression levels.

First, we note that ratio $\left\|\boldsymbol{p}_j\right\|_1 / \left\|\boldsymbol{p}_j\right\|_2$ is independent of the magnitude of the expression vector $\boldsymbol{p}_j$. In other words, normalization does not affect the ratio;

$$\frac{\left\|\dfrac{\boldsymbol{p}_j}{\left\|\boldsymbol{p}_j\right\|_2}\right\|_1}{\left\|\dfrac{\boldsymbol{p}_j}{\left\|\boldsymbol{p}_j\right\|_2}\right\|_2} = \frac{\displaystyle\sum_{i=1}^{m} \frac{|p_{ij}|}{\left\|\boldsymbol{p}_j\right\|_2}}{\sqrt{\displaystyle\sum_{i=1}^{m} \frac{(p_{ij})^2}{\left\|\boldsymbol{p}_j\right\|_2^2}}} = \frac{\dfrac{\left\|\boldsymbol{p}_j\right\|_1}{\left\|\boldsymbol{p}_j\right\|_2}}{\dfrac{\left\|\boldsymbol{p}_j\right\|_2}{\left\|\boldsymbol{p}_j\right\|_2}} = \frac{\left\|\boldsymbol{p}_j\right\|_1}{\left\|\boldsymbol{p}_j\right\|_2}. \tag{1.16}$$

On the basis of this, we only consider normalized expression vectors $\boldsymbol{p}_j / \left\|\boldsymbol{p}_j\right\|_2$ without loss of generality.

By definition, L^2 norm of a normalized expression vector (the denominator of the most left-hand side of (1.16)) equals to one. Thus, the ratio we are considering equals to the L^1 norm of the normalized expression vector;

$$\frac{\left\|\boldsymbol{p}_j\right\|_1}{\left\|\boldsymbol{p}_j\right\|_2} = \left\|\frac{\boldsymbol{p}_j}{\left\|\boldsymbol{p}_j\right\|_2}\right\|_1. \tag{1.17}$$

Here, we write

$$\frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} = \begin{pmatrix} \tilde{p}_{1j} \\ \tilde{p}_{2j} \\ \vdots \\ \tilde{p}_{mj} \end{pmatrix}, \tag{1.18}$$

where $\tilde{p}_{1j}, \tilde{p}_{2j}, \ldots, \tilde{p}_{mj} \geq 0$. Then,

$$\frac{\|\boldsymbol{p}_j\|_1}{\|\boldsymbol{p}_j\|_2} = \left\| \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \right\|_1 = \sum_{i=1}^{m} \tilde{p}_{ij}. \tag{1.19}$$

Note that $\sum_{i=1}^{m} (\tilde{p}_{ij})^2 = 1$ holds because of normalization. Therefore, any normalized expression vector corresponds to a point on the first orthant division of the unit $(m-1)$-sphere $\sum_{i=1}^{m} (x_i)^2 = 1$ $(x_1, \ldots, x_m \geq 0)$.

Next we consider a hyperplane $\sum_{i=1}^{m} x_i = k$ which passes through the point (1.18). Since all the coefficients of this hyperplane are equal, all the m intercepts are also equal. The intercept value is $k = \sum_{i=1}^{m} \tilde{p}_{ij}$, thus equals to the ratio of (1.19). In other words, the ratio (1.19) appears as an intercept of the hyperplane passing through the point corresponding to the normalized vector $\boldsymbol{p}_j / \|\boldsymbol{p}_j\|_2$ with all the coefficients equal to one.

By simple calculation, one can see that the two surfaces $\sum_{i=1}^{m} (x_i)^2 = 1$ $(x_1, \ldots, x_m \geq 0)$ and $\sum_{i=1}^{m} x_i = k$ intersect when $1 \leq k \leq \sqrt{m}$. [1] The intercept value $k = \sum_{i=1}^{m} \tilde{p}_{ij}$ takes the maximum $k = \sqrt{m}$ when the normalized expression vector points to the "center" of

[1] In other words, $\|\boldsymbol{p}_j\|_2 \leq \|\boldsymbol{p}_j\|_1 \leq \sqrt{m}\|\boldsymbol{p}_j\|_2$ holds.

the first orthant division of the unit $(m-1)$-sphere, i.e., when

$$\frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} = \begin{pmatrix} \tilde{p}_{1j} \\ \tilde{p}_{2j} \\ \vdots \\ \tilde{p}_{mj} \end{pmatrix} = \frac{1}{\sqrt{m}} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{1.20}$$

This means that the expression level is even and constant across the conditions. When this evenness of expression level breaks, the intercept value k decreases, and it attains the minimum $k = 1$ when the normalized expression vector overlaps with an axis, i.e., when

$$\frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} = \begin{pmatrix} \tilde{p}_{1j} \\ \vdots \\ \tilde{p}_{(\mu-1)j} \\ \tilde{p}_{\mu j} \\ \tilde{p}_{(\mu+1)j} \\ \vdots \\ \tilde{p}_{mj} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{1.21}$$

which corresponds to a completely "condition-specific expression pattern" (μ is the condition's index).

See Figures S8A and S8B for a graphical explanation of the argument for the two- and three-dimensional cases.

1.10 Proteome structure obtained with PCA

As mentioned in the main text, we confirmed that PCA could find a proteome structure (Figure S6H) similar to the cosine similarity LE structure (Figures 5L and 6D). Since cosine similarity of expression vectors is inner product of the L^2-normalized expression vectors, we also performed L^2 normalization of proteome data before applying PCA in this PCA analysis. In other words, we applied PCA to a normalized proteome data $\begin{bmatrix} \boldsymbol{p}_1/\|\boldsymbol{p}_1\|_2 & \cdots & \boldsymbol{p}_n/\|\boldsymbol{p}_n\|_2 \end{bmatrix}$ (See 2.1.5).

We remark that, despite the structural similarity between the PCA structure and the cosine similarity LE structure, cosine similarity LE has an advantage over PCA in that the relative proximity of positions reflects the strength of stoichiometry conservation between each element. In addition, as shown in the main text and the section 2, cosine similarity LE of omics data has a direct quantitative connection to cellular Raman spectra, which is not the case for PCA.

2 Mathematical analysis and details

To clarify what is non-trivial in the correspondence between LDA Raman and cosine similarity LE proteome, here we derive rigorous mathematical relations for the correspondence through linear algebraic calculation.

2.1 Mathematics behind the correspondence

2.1.1 Notations

We use following notations:

- $\mathbf{1}_x$ denotes an x-dimensional column vector of ones, and $\mathbf{0}_x$ does an x-dimensional zero column vector.
- A vector without a hat (e.g. $\boldsymbol{x}$) is a column vector, and a vector with a hat (e.g. $\hat{\boldsymbol{x}}$) is a row vector.
- I denotes an identity matrix, and O does a zero matrix.
- For a square matrix $X = (x_{ij})$, $\mathrm{diag}(X) = (\delta_{ij} x_{ij})$, and for a vector $\boldsymbol{x} = (x_i)$, $\mathrm{diag}(\boldsymbol{x}) = (\delta_{ij} x_i)$, where δ_{ij} is the Kronecker delta.
- For a matrix X, $X[i,:]$ denotes the i-th row of X, and $X[:,j]$ denotes the j-th column of X.

2.1.2 Preparations

Let l be the number of cells in each condition, m be the number of conditions, n be the original dimension of proteome, i.e. the number of protein species in the proteome

data, and s' be the original dimension of Raman spectra after the application of the Savitzky-Golay filter. In our main data, $l = 38, m = 15, n = 2058$, and $s' = 599$.

Raman Spectra

Original preprocessed Raman data Let

$$\hat{\boldsymbol{x}}_j'^{(i)} = \begin{pmatrix} x_{j1}'^{(i)} & \cdots & x_{js'}'^{(i)} \end{pmatrix} \tag{2.1}$$

be a preprocessed Raman spectrum from cell j under condition i (See 1.1.3). The prime $'$ denotes that the variable is the original preprocessed data. The $\hat{\boldsymbol{x}}_j'^{(i)}$ $(1 \leq j \leq l)$ are collected in an $l \times s'$ matrix:

$$X_i' = \begin{bmatrix} \hat{\boldsymbol{x}}_1'^{(i)} \\ \vdots \\ \hat{\boldsymbol{x}}_l'^{(i)} \end{bmatrix} = \left(x_{jk}'^{(i)} \right)_{1 \leq j \leq l, 1 \leq k \leq s'}. \tag{2.2}$$

Combining X_i' from different conditions, one can define an $(lm) \times s'$ matrix

$$X' = \begin{bmatrix} X_1' \\ \vdots \\ X_m' \end{bmatrix}, \tag{2.3}$$

which contains all the preprocessed Raman data.

PCA and centering Before LDA, principal component analysis (PCA) was first applied to the preprocessed Raman data to reduce noise. The covariance matrix is

$$C_{X'} \coloneqq \frac{1}{lm-1}\left(X' - \frac{1}{lm}\mathbf{1}_{lm}(\mathbf{1}_{lm})^\top X'\right)^\top \left(X' - \frac{1}{lm}\mathbf{1}_{lm}(\mathbf{1}_{lm})^\top X'\right), \tag{2.4}$$

which is positive semi-definite. PCA is formulated as the following eigenvalue problem of $C_{X'}$:

$$C_{X'}V_{\mathrm{PCA}} = V_{\mathrm{PCA}}\Lambda_{\mathrm{PCA}}, \tag{2.5}$$

where Λ_{PCA} is a diagonal matrix with the eigenvalues of $C_{X'}$ as its diagonal elements in decreasing order from the upper left, and V_{PCA} is an orthogonal matrix consisting of the eigenvectors of $C_{X'}$ as its columns. Using the first s ($1 \leq s \leq s'$) columns of $V_{\mathrm{PCA}} = \begin{bmatrix}\boldsymbol{w}_1 & \cdots & \boldsymbol{w}_{s'}\end{bmatrix}$, i.e. the columns corresponding the first s largest eigenvalues, we obtain an $(lm) \times s$ matrix representing the post PCA data

$$X \coloneqq \left(X' - \frac{1}{lm}\mathbf{1}_{lm}(\mathbf{1}_{lm})^\top X'\right)\begin{bmatrix}\boldsymbol{w}_1 & \cdots & \boldsymbol{w}_s\end{bmatrix}. \tag{2.6}$$

Here, $\boldsymbol{w}_k$ is the k-th PCA coefficient vector, and s is the reduced dimension of the Raman spectra. The subtraction of the $\frac{1}{lm}\mathbf{1}_{lm}(\mathbf{1}_{lm})^\top X'$ is for centering the data. In our case, the top 218 principal components (i.e., $s = 218$) explaining 98% of the variance were used to reduce noise and dimensionality.

Let $\hat{\boldsymbol{x}}_j^{(i)}$ be the post PCA Raman spectrum from cell j under condition i, i.e.,

$$\hat{\boldsymbol{x}}_j^{(i)} = \begin{pmatrix}x_{j1}^{(i)} & \cdots & x_{js}^{(i)}\end{pmatrix}, \tag{2.7}$$

and X_i be the collection of $\hat{\boldsymbol{x}}_j^{(i)}$ $(1 \leq j \leq l)$; i.e.,

$$X_i = \begin{bmatrix} \hat{\boldsymbol{x}}_1^{(i)} \\ \vdots \\ \hat{\boldsymbol{x}}_l^{(i)} \end{bmatrix} = \left(x_{jk}^{(i)} \right)_{1 \leq j \leq l, 1 \leq k \leq s}. \tag{2.8}$$

Then, X is written as

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}. \tag{2.9}$$

From (2.6),

$$(\mathbf{1}_{lm})^\top X = \mathbf{0}, \tag{2.10}$$

namely, for any k,

$$\sum_{i=1}^{m} \sum_{j=1}^{l} x_{jk}^{(i)} = 0, \tag{2.11}$$

which means that the post PCA data is centered.

<u>Population average in each condition</u> Let $\hat{\boldsymbol{x}}_i$ be the population average of the post PCA spectra of cells under condition i. Then,

$$\hat{\boldsymbol{x}}_i = \frac{1}{l} (\mathbf{1}_l)^\top X_i. \tag{2.12}$$

Also,

$$\hat{\boldsymbol{x}}_i = \begin{pmatrix} \bar{x}_{i1} & \cdots & \bar{x}_{is} \end{pmatrix}, \tag{2.13}$$

where

$$\bar{x}_{ik} = \frac{1}{l}\sum_{j=1}^{l} x_{jk}^{(i)}. \tag{2.14}$$

We define an $m \times s$ matrix

$$\bar{X} = \begin{bmatrix} \hat{\boldsymbol{x}}_1 \\ \vdots \\ \hat{\boldsymbol{x}}_m \end{bmatrix} = (\bar{x}_{ik})_{1\le i\le m, 1\le k\le s}\,. \tag{2.15}$$

Each row of $\bar{X}$ corresponds to a condition. From (2.11) and (2.14), for any k,

$$\sum_{i=1}^{m} \bar{x}_{ik} = 0. \tag{2.16}$$

Namely,

$$(\mathbf{1}_m)^\top \bar{X} = \mathbf{0}. \tag{2.17}$$

These relations mean that $\bar{X}$ is also centered.

<u>LDA</u> The within-class covariance matrix is

$$\begin{aligned} C_I &:= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{l-1}\left(X_i - \frac{1}{l}\mathbf{1}_l\left(\mathbf{1}_l\right)^\top X_i\right)^\top\left(X_i - \frac{1}{l}\mathbf{1}_l\left(\mathbf{1}_l\right)^\top X_i\right) && (2.18)\\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{l-1}\left(X_i - \mathbf{1}_l\hat{\boldsymbol{x}}_i\right)^\top\left(X_i - \mathbf{1}_l\hat{\boldsymbol{x}}_i\right). && (2.19)\end{aligned}$$

Here, (2.12) was used. The between-class covariance matrix is

$$\begin{aligned} C_B &:= \frac{1}{m-1}\left(\bar{X} - \frac{1}{m}\mathbf{1}_m\left(\mathbf{1}_m\right)^\top \bar{X}\right)^\top\left(\bar{X} - \frac{1}{m}\mathbf{1}_m\left(\mathbf{1}_m\right)^\top \bar{X}\right) && (2.20)\\ &= \frac{1}{m-1}\bar{X}^\top\bar{X}. && (2.21)\end{aligned}$$

Here, (2.17) was used. Assume $\mathrm{rank}(C_I) = s$ and $\mathrm{rank}(C_B) = m - 1$ (the maximum possible values). In fact, $\mathrm{rank}(C_I) = s$ and $\mathrm{rank}(C_B) = m - 1$ in our data. Note that $s > m - 1$. From the definitions of C_I and C_B above, both are positive semi-definite. LDA is formulated as the following generalized eigenvalue problem:

$$C_B V_{\mathrm{LDA}} = C_I V_{\mathrm{LDA}} \Lambda'_{\mathrm{LDA}}, \tag{2.22}$$

where Λ'_{LDA} is a diagonal matrix, and V_{LDA} is an $s \times (m-1)$ matrix that simultaneously diagonize C_B and C_I (to Λ_{LDA}):

$$(V_{\mathrm{LDA}})^\top C_B V_{\mathrm{LDA}} = (V_{\mathrm{LDA}})^\top C_I V_{\mathrm{LDA}} \Lambda'_{\mathrm{LDA}} = \Lambda_{\mathrm{LDA}}. \tag{2.23}$$

Here, the diagonal elements in Λ'_{LDA} were in decreasing order from the upper left. In our analysis, the columns of V_{LDA} were normalized.

Using V_{LDA}, we obtain an $m \times (m-1)$ matrix representing the post LDA data

$$R \coloneqq \bar{X} V_{\mathrm{LDA}}. \tag{2.24}$$

Each row of R represents a dimension-reduced Raman spectrum of each condition. Let us write the h-th $(1 \leq h \leq m-1)$ column of R as

$$\boldsymbol{r}_h \coloneqq \begin{pmatrix} r_{1h} \\ \vdots \\ r_{mh} \end{pmatrix}. \tag{2.25}$$

Then,

$$R = \begin{bmatrix} \boldsymbol{r}_1 & \cdots & \boldsymbol{r}_{m-1} \end{bmatrix} = (r_{ih})_{1 \le i \le m, 1 \le h \le m-1}. \tag{2.26}$$

Transforming $R^\top R$ gives

$$\begin{aligned} R^\top R &= (V_{\mathrm{LDA}})^\top \bar{X}^\top \bar{X} V_{\mathrm{LDA}} && (2.27) \\ &= (m-1)(V_{\mathrm{LDA}})^\top C_B V_{\mathrm{LDA}} \quad (\because (2.20)) && (2.28) \\ &= (m-1)\Lambda_{\mathrm{LDA}} \quad (\because (2.23)). && (2.29) \end{aligned}$$

Therefore, $R^\top R$ is a diagonal matrix, and $\boldsymbol{r}_h$ are orthogonal to each other. As all the diagonal elements of the diagonal matrix Λ_{LDA} is positive, $\mathrm{rank}(R) = m-1$. Furthermore,

$$\begin{aligned} (\mathbf{1}_m)^\top R &= (\mathbf{1}_m)^\top \bar{X} V_{\mathrm{LDA}} && (2.30) \\ &= (\mathbf{0}_{m-1})^\top \quad (\because (2.17)). && (2.31) \end{aligned}$$

Namely, for any h,

$$\sum_{i=1}^{m} r_{ih} = 0. \tag{2.32}$$

This means that, as the data is centered, all the columns of R is perpendicular to $\mathbf{1}_m$.

Proteome

Let

$$\boldsymbol{p}_j = \begin{pmatrix} p_{1j} \\ \vdots \\ p_{mj} \end{pmatrix} \tag{2.33}$$

be the absolute abundances of the j-th protein. $\boldsymbol{p}_j$ are collected in an $m \times n$ matrix:

$$P := \begin{bmatrix} \boldsymbol{p}_1 & \cdots & \boldsymbol{p}_n \end{bmatrix} = (p_{ij})_{1 \leq i \leq m, 1 \leq j \leq n} \,. \tag{2.34}$$

p_{ij} is the absolute abundance of the j-th protein in condition i.

We assume $\operatorname{rank}(P) = m$, i.e., proteome vectors for different conditions are linear independent. Actually, $\operatorname{rank}(P) = m$ in our data. We also assume that proteins with zero expression in all the m conditions had been excluded from the proteome data.

2.1.3 Linear transformation between LDA Raman and proteome

We define an $m \times m$ matrix

$$R_E \coloneqq \begin{bmatrix} \mathbf{1}_m & R \end{bmatrix} = \begin{bmatrix} \mathbf{1}_m & \boldsymbol{r}_1 & \cdots & \boldsymbol{r}_{m-1} \end{bmatrix}. \tag{2.35}$$

We denote the first column of R_E as $\boldsymbol{r}_0$. Hence,

$$\boldsymbol{r}_0 = \mathbf{1}_m. \tag{2.36}$$

From (2.29) and (2.31),

$$(R_E)^\top R_E = \begin{bmatrix} (\mathbf{1}_m)^\top \\ R^\top \end{bmatrix} \begin{bmatrix} \mathbf{1}_m & R \end{bmatrix} = \begin{bmatrix} m & \mathbf{0} \\ \mathbf{0} & R^\top R \end{bmatrix} = \begin{bmatrix} m & \mathbf{0} \\ \mathbf{0} & (m-1)\Lambda_{\mathrm{LDA}} \end{bmatrix}. \tag{2.37}$$

Therefore, $(R_E)^\top R_E$ is also a diagonal matrix, and R_E has full rank. For convenience, we write

$$(R_E)^\top R_E = \Lambda_{R_E} = \begin{pmatrix} m & & & \\ & (m-1)\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & (m-1)\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix}. \tag{2.38}$$

Here, we consider singular value decomposition (SVD) of R_E; i.e.,

$$R_E = U_{R_E} \Sigma_{R_E} (V_{R_E})^\top, \tag{2.39}$$

where Σ_{R_E} is a diagonal matrix whose diagonal elements are the singular values of R_E, and $(U_{R_E})^\top U_{R_E} = (V_{R_E})^\top V_{R_E} = I$. [2] As Λ_{R_E} is the eigenvalue matrix of $(R_E)^\top R_E$,

$$(R_E)^\top R_E = \Lambda_{R_E} = (\Sigma_{R_E})^2 , \tag{2.44}$$

and

$$\Sigma_{R_E} = \begin{pmatrix} \sqrt{m} & & & \\ & \sqrt{(m-1)\lambda_{\mathrm{LDA1}}} & & \\ & & \ddots & \\ & & & \sqrt{(m-1)\lambda_{\mathrm{LDA}(m-1)}} \end{pmatrix}. \tag{2.45}$$

Now, we consider linear transformation between P and R_E. We introduce the $n \times m$

[2] We can set $V_{R_E} = I$ in the following way. Let

$$U_{R_E} \coloneqq R_E \left(\Lambda_{R_E}\right)^{-1/2} . \tag{2.40}$$

Then,

$$\begin{aligned} (U_{R_E})^\top U_{R_E} &= I, & (2.41) \\ (U_{R_E})^\top R_E &= (\Lambda_{R_E})^{1/2} . & (2.42) \end{aligned}$$

Thus, SVD of R_E can be written as

$$R_E = U_{R_E} \left(\Lambda_{R_E}\right)^{1/2} I. \tag{2.43}$$

coefficient matrix $B_E = \begin{bmatrix} \boldsymbol{b}_0 & \cdots & \boldsymbol{b}_{m-1} \end{bmatrix}$ that connects P and R_E as

$$\begin{aligned} P &= R_E \left(B_E\right)^\top && (2.46) \\ &= \begin{bmatrix} \boldsymbol{r}_0 & \boldsymbol{r}_1 & \cdots & \boldsymbol{r}_{m-1} \end{bmatrix} \begin{bmatrix} \left(\boldsymbol{b}_0\right)^\top \\ \left(\boldsymbol{b}_1\right)^\top \\ \vdots \\ \left(\boldsymbol{b}_{m-1}\right)^\top \end{bmatrix} && (2.47) \\ &= \mathbf{1}_m \left(\boldsymbol{b}_0\right)^\top + \boldsymbol{r}_1 \left(\boldsymbol{b}_1\right)^\top + \cdots + \boldsymbol{r}_{m-1} \left(\boldsymbol{b}_{m-1}\right)^\top . && (2.48) \end{aligned}$$

R_E has full rank and is therefore invertible. Thus, B_E is obtained by

$$B_E \coloneqq P^\top \left(\left(R_E\right)^{-1}\right)^\top . \tag{2.49}$$

From the viewpoint of linear regression, $\boldsymbol{b}_0$ can be regarded as the constant terms and $\boldsymbol{b}_h$ $(1 \leq h \leq m-1)$ is the coefficients for the h-th LDA dimension.

We can rewrite P using the row vectors in R_E and B_E. Writing the i-th $(1 \leq i \leq m)$ row of R_E as $\hat{\boldsymbol{r}}_i = \begin{pmatrix} 1 & r_{i1} & \cdots & r_{i(m-1)} \end{pmatrix}$,

$$R_E = \begin{bmatrix} \hat{\boldsymbol{r}}_1 \\ \vdots \\ \hat{\boldsymbol{r}}_m \end{bmatrix} . \tag{2.50}$$

Likewise, writing the j-th ($1 \leq j \leq n$) row of B_E as $\hat{\boldsymbol{b}}_j = \begin{pmatrix} b_{j0} & \cdots & b_{j(m-1)} \end{pmatrix}$,

$$B_E = \begin{bmatrix} \hat{\boldsymbol{b}}_1 \\ \vdots \\ \hat{\boldsymbol{b}}_n \end{bmatrix}. \tag{2.51}$$

Then, (2.46) can be written in another way;

$$\begin{aligned} P &= R_E \left(B_E\right)^\top \\ &= \begin{bmatrix} \hat{\boldsymbol{r}}_1 \\ \vdots \\ \hat{\boldsymbol{r}}_m \end{bmatrix} \begin{bmatrix} \left(\hat{\boldsymbol{b}}_1\right)^\top & \cdots & \left(\hat{\boldsymbol{b}}_n\right)^\top \end{bmatrix} = \begin{pmatrix} \hat{\boldsymbol{r}}_1 \left(\hat{\boldsymbol{b}}_1\right)^\top & \cdots & \hat{\boldsymbol{r}}_1 \left(\hat{\boldsymbol{b}}_n\right)^\top \\ \vdots & \ddots & \vdots \\ \hat{\boldsymbol{r}}_m \left(\hat{\boldsymbol{b}}_1\right)^\top & \cdots & \hat{\boldsymbol{r}}_m \left(\hat{\boldsymbol{b}}_n\right)^\top \end{pmatrix}. \end{aligned} \tag{2.52}$$

The interpretation of each vector is summarized in Table S9.

Table S9: **Interpretations of $\boldsymbol{r}_h$, $\hat{\boldsymbol{r}}_i$, $\boldsymbol{b}_h$, and $\hat{\boldsymbol{b}}_j$**

Matrix		Vector	Dimension	Description
R_E	column	$\boldsymbol{r}_h$ ($h = 0, \ldots, m-1$)	m	list of h-th LDA coordinates of mean LDA Raman of all the conditions
	row	$\hat{\boldsymbol{r}}_i$ ($i = 1, \ldots, m$)	m	mean LDA Raman of condition i
B_E	column	$\boldsymbol{b}_h$ ($h = 0, \ldots, m-1$)	n	list of coefficients of all the proteins for the h-th LDA axis
	row	$\hat{\boldsymbol{b}}_j$ ($j = 1, \ldots, n$)	m	coefficients for protein j

2.1.4 Relation between LDA Raman and Ω_B

Here, we discuss the spatial correspondence between the Raman distribution in LDA space and the normalized Raman-proteome coefficient proteome structure (Figures 6B and 6C).

Connecting LDA Raman and Raman-omics transformation coefficients

Let us consider $(R_E)^\top R_E (B_E)^\top$ in two ways. In the first approach,

$$
\begin{aligned}
(R_E)^\top \left(R_E (B_E)^\top\right) &= (R_E)^\top P && (2.53)\\
&= \begin{bmatrix} (\hat{\boldsymbol{r}}_1)^\top & \cdots & (\hat{\boldsymbol{r}}_m)^\top \end{bmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{pmatrix} && (2.54)\\
&= \begin{bmatrix} \sum_{i=1}^m p_{i1} (\hat{\boldsymbol{r}}_i)^\top & \cdots & \sum_{i=1}^m p_{in} (\hat{\boldsymbol{r}}_i)^\top \end{bmatrix}. && (2.55)
\end{aligned}
$$

In the second approach,

$$
\begin{aligned}
&\left((R_E)^\top R_E\right)(B_E)^\top \\
&= (\Sigma_{R_E})^2 (B_E)^\top && (2.56)\\
&= \begin{pmatrix} m & & & \\ & (m-1)\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & (m-1)\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix} \begin{bmatrix} \left(\hat{\boldsymbol{b}}_1\right)^\top & \cdots & \left(\hat{\boldsymbol{b}}_n\right)^\top \end{bmatrix}. && (2.57)
\end{aligned}
$$

Comparing the two calculations yields

$$
(R_E)^\top P = (\Sigma_{R_E})^2 (B_E)^\top . \qquad (2.58)
$$

This is equivalent to

$$\sum_{i=1}^{m} p_{ij} \left(\hat{\boldsymbol{r}}_i\right)^\top = \begin{pmatrix} m & & & \\ & (m-1)\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & (m-1)\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix} \left(\hat{\boldsymbol{b}}_j\right)^\top \tag{2.59}$$

for any protein j.

Normalization with constant terms

Next we consider the normalization of the matrices in (2.58) with constant terms. We first define the normalized coefficient matrix as

$$B_E^{\mathrm{norm}} := \begin{pmatrix} 1 & \frac{b_{11}}{b_{10}} & \dots & \frac{b_{1(m-1)}}{b_{10}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{b_{n1}}{b_{n0}} & \dots & \frac{b_{n(m-1)}}{b_{n0}} \end{pmatrix} = \mathrm{diag}\left(\boldsymbol{b}_0\right)^{-1} B_E. \tag{2.60}$$

This is normalization of the coefficients by the constant terms. Furthermore, we normalize Σ_{R_E} as

$$\Sigma_{R_E}^{\mathrm{norm}} := \frac{1}{\left(\Sigma_{R_E}\right)_{11}} \Sigma_{R_E} = \frac{1}{\sqrt{m}} \Sigma_{R_E} \tag{2.61}$$

$$= \begin{pmatrix} 1 & & & \\ & \sqrt{\frac{m-1}{m}\lambda_{\mathrm{LDA1}}} & & \\ & & \ddots & \\ & & & \sqrt{\frac{m-1}{m}\lambda_{\mathrm{LDA}(m-1)}} \end{pmatrix}. \tag{2.62}$$

Thus, the right-hand side of (2.58) is

$$(\Sigma_{R_E})^2 (B_E)^\top = m \left(\Sigma_{R_E}^{\text{norm}}\right)^2 (B_E^{\text{norm}})^\top \operatorname{diag}(\boldsymbol{b}_0). \tag{2.63}$$

Since the first row of $(R_E)^\top$ is $(\mathbf{1}_m)^\top$, one can rewrite the left-hand side of (2.58) as

$$\begin{aligned}
(R_E)^\top P &= \begin{bmatrix} \frac{\sum_{i=1}^m p_{i1}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{i1}} & \dots & \frac{\sum_{i=1}^m p_{in}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{in}} \end{bmatrix} \begin{pmatrix} \sum_{i=1}^m p_{i1} & & \\ & \ddots & \\ & & \sum_{i=1}^m p_{in} \end{pmatrix} && (2.64) \\
&= \begin{bmatrix} \frac{\sum_{i=1}^m p_{i1}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{i1}} & \dots & \frac{\sum_{i=1}^m p_{in}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{in}} \end{bmatrix} \operatorname{diag}\left((\mathbf{1}_m)^\top P\right). && (2.65)
\end{aligned}$$

Here,

$$\begin{aligned}
(\mathbf{1}_m)^\top P &= (\mathbf{1}_m)^\top R_E (B_E)^\top && (2.66) \\
&= \begin{bmatrix} m & (\mathbf{0}_{m-1})^\top \end{bmatrix} (B_E)^\top && (2.67) \\
&= m (\boldsymbol{b}_0)^\top. && (2.68)
\end{aligned}$$

Therefore,

$$\boldsymbol{b}_0 = \frac{1}{m} \begin{pmatrix} (\mathbf{1}_m)^\top \boldsymbol{p}_1 \\ \vdots \\ (\mathbf{1}_m)^\top \boldsymbol{p}_n \end{pmatrix} = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^m p_{i1} \\ \vdots \\ \frac{1}{m} \sum_{i=1}^m p_{in} \end{pmatrix}. \tag{2.69}$$

This relation indicates that the constant term for each protein is its average abundance. Consequently,

$$(R_E)^\top P = m \begin{bmatrix} \frac{\sum_{i=1}^m p_{i1}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{i1}} & \dots & \frac{\sum_{i=1}^m p_{in}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^m p_{in}} \end{bmatrix} \operatorname{diag}(\boldsymbol{b}_0). \tag{2.70}$$

Therefore, we obtain

$$\begin{bmatrix} \frac{\sum_{i=1}^{m} p_{i1}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^{m} p_{i1}} & \cdots & \frac{\sum_{i=1}^{m} p_{in}(\hat{\boldsymbol{r}}_i)^\top}{\sum_{i=1}^{m} p_{in}} \end{bmatrix} = \left(\Sigma_{R_E}^{\mathrm{norm}}\right)^2 \left(B_E^{\mathrm{norm}}\right)^\top . \tag{2.71}$$

Equivalently,

$$\frac{\sum_{i=1}^{m} p_{ij} \left(\hat{\boldsymbol{r}}_i\right)^\top}{\sum_{i=1}^{m} p_{ij}} = \begin{pmatrix} 1 & & & \\ & \frac{m-1}{m}\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & \frac{m-1}{m}\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{b_{j1}}{b_{j0}} \\ \vdots \\ \frac{b_{j(m-1)}}{b_{j0}} \end{pmatrix} . \tag{2.72}$$

for any protein j. This means that the normalized coefficients of each protein are mainly determined by the weighted averages of the Raman vectors, where the weight is the abundances of the protein.

As we already saw in (1.6) or (1.7), this equation also shows that protein pairs whose abundance ratio remains constant over all the conditions have identical normalized coefficients.

Special case — condition-specific protein

Consider an imaginary condition-specific protein γ whose abundance is c (> 0) under condition Γ and zero under the other conditions; i.e.,

$$p_{i\gamma} = \begin{cases} 0 & (i \neq \Gamma) \\ c & (i = \Gamma) \end{cases} . \tag{2.73}$$

From (2.59),

$$\begin{pmatrix} c & & & \\ & c & & \\ & & \ddots & \\ & & & c \end{pmatrix} (\hat{\boldsymbol{r}}_\Gamma)^\top = \begin{pmatrix} m & & & \\ & (m-1)\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & (m-1)\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix} \left(\hat{\boldsymbol{b}}_\gamma\right)^\top, \quad (2.74)$$

which indicates that the LDA Raman of condition Γ, $(\hat{\boldsymbol{r}}_\Gamma)^\top$, and Raman-proteome coefficients for Γ-specific protein γ, $\left(\hat{\boldsymbol{b}}_\gamma\right)^\top$, are in the same orthant. The normalized version is obtained by dividing the both sides by the first row (or from (2.72));

$$(\hat{\boldsymbol{r}}_\Gamma)^\top = \begin{pmatrix} 1 & & & \\ & \frac{m-1}{m}\lambda_{\mathrm{LDA1}} & & \\ & & \ddots & \\ & & & \frac{m-1}{m}\lambda_{\mathrm{LDA}(m-1)} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{b_{\gamma 1}}{b_{\gamma 0}} \\ \vdots \\ \frac{b_{\gamma(m-1)}}{b_{\gamma 0}} \end{pmatrix} \quad (2.75)$$

$$= \left(\Sigma_{R_E}^{\mathrm{norm}}\right)^2 \frac{\left(\hat{\boldsymbol{b}}_\gamma\right)^\top}{b_{\gamma 0}}, \quad (2.76)$$

which shows that the LDA Raman of condition Γ, $(\hat{\boldsymbol{r}}_\Gamma)^\top$, and the normalized Raman-proteome coefficients of Γ-specific protein γ, $\left(\hat{\boldsymbol{b}}_\gamma\right)^\top / b_{\gamma 0}$, are in the same orthant.

Application to main data

The LDA Raman distribution shown in Figure 6B corresponds to $\hat{\boldsymbol{r}}_i$ (scatter plots between different columns of R_E). On the other hand, the normalized coefficient proteome structure in Figure 6C is the scatter plots between different columns of B_E^{norm}. The above linear algebra explains the correspondence between the two. In addition, from (2.72), one

can understand that the homeostatic core distributes around the center of the structure because its member proteins are expressed in all the conditions.

The relation (2.76) was obtained by considering an imaginary protein whose expression levels were zero under all conditions except for one condition. To confirm this relation with actual data, we picked an almost-condition-specific protein (PaaE, highly expressed in LB condition) and a non-condition-specific protein (AcrR), and confirmed that the former approximately satisfied (2.76), while the relation did not hold for the latter (Figure S10).

2.1.5 Relation between Ω_B and Ω_{LE}

Here, we discuss the spatial correspondence between the normalized Raman-proteome coefficient proteome structure and the cosine similarity LE proteome structure (Figures 6C and 6D).

Cosine similarity LE proteome structure

Consider an undirected graph where each node corresponds to one type of protein. As previously explained, let the graph be a complete graph, namely every pair of nodes is connected by an edge. Each edge is weighted with cosine similarity between the two types of protein connected by the edge. Cosine similarity of protein i and protein j is given by

$$\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} \coloneqq \frac{\boldsymbol{p}_i \cdot \boldsymbol{p}_j}{\|\boldsymbol{p}_i\|_2 \left\|\boldsymbol{p}_j\right\|_2}, \tag{2.77}$$

where $\boldsymbol{p}_i \cdot \boldsymbol{p}_j \left(= (\boldsymbol{p}_i)^\top \boldsymbol{p}_j\right)$ is the inner product of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, and $\|\boldsymbol{p}_i\|_2 \left(= \sqrt{(\boldsymbol{p}_i)^\top \boldsymbol{p}_i}\right)$ is the L^2-norm (Euclidean norm) of $\boldsymbol{p}_i$. Cosine similarity $\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}$ evaluates how similar the expression patterns are between protein i and protein j. When the abundance ratio between protein i and protein j remains constant over all the m conditions, $\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}$ takes the maximum value 1.

The adjacency matrix of this graph is given by

$$A \coloneqq \left(\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}\right)_{ij} = \left(\frac{\boldsymbol{p}_i \cdot \boldsymbol{p}_j}{\|\boldsymbol{p}_i\|_2 \left\|\boldsymbol{p}_j\right\|_2}\right)_{ij}, \tag{2.78}$$

and the degree matrix of this graph is

$$D = \mathrm{diag}\left(A\mathbf{1}_n\right). \tag{2.79}$$

For simplicity, diagonal element (i,i) of A is $\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_i} = 1$ for any protein i, i.e., each node has a loop. A is $n \times n$ and real symmetric and D is $n \times n$ and diagonal. Then, the Laplacian matrix is given by

$$L \coloneqq D - A, \tag{2.80}$$

which is an $n \times n$ symmetric matrix, and the symmetric normalized Laplacian is given by

$$L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}, \tag{2.81}$$

which is an $n \times n$ symmetric matrix.

Here we define $\hat{P}$ by normalizing the columns of $P = \begin{bmatrix} \boldsymbol{p}_1 & \cdots & \boldsymbol{p}_n \end{bmatrix}$;

$$\hat{P} \coloneqq \begin{bmatrix} \frac{\boldsymbol{p}_1}{\|\boldsymbol{p}_1\|_2} & \cdots & \frac{\boldsymbol{p}_n}{\|\boldsymbol{p}_n\|_2} \end{bmatrix} = P \operatorname{diag}\left(P^\top P\right)^{-1/2}. \tag{2.82}$$

By using $\hat{P}$, A is rewritten as

$$\begin{aligned} A = \left(\cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}\right)_{ij} &= \left(\frac{\boldsymbol{p}_i \cdot \boldsymbol{p}_j}{\|\boldsymbol{p}_i\|_2 \|\boldsymbol{p}_j\|_2}\right)_{ij} && (2.83) \\ &= \left(P \operatorname{diag}\left(P^\top P\right)^{-1/2}\right)^\top P \operatorname{diag}\left(P^\top P\right)^{-1/2} && (2.84) \\ &= \operatorname{diag}\left(P^\top P\right)^{-1/2} P^\top P \operatorname{diag}\left(P^\top P\right)^{-1/2} && (2.85) \\ &= \hat{P}^\top \hat{P}. && (2.86) \end{aligned}$$

Consider an eigenproblem

$$L_{\mathrm{sym}} V_{\mathrm{sym}} = V_{\mathrm{sym}} \Lambda_{\mathrm{LE}}, \tag{2.87}$$

where Λ is an $n \times n$ diagonal matrix in which the eigenvalues of L_{sym} are arranged in

increasing order from the upperleft, and columns of V_{sym} are the normalized eigenvectors of L_{sym} corresponding to the eigenvalues. Denote

$$\Lambda_{\mathrm{LE}} = \begin{pmatrix} \lambda_{\mathrm{LE}0} & & \\ & \ddots & \\ & & \lambda_{\mathrm{LE}(n-1)} \end{pmatrix}. \tag{2.88}$$

Here, L_{sym} has the following four characteristics:

- L_{sym} is positive semi-definite. See, for example, [20] for the proof.

- In an undirected graph with non-negative weights, the number of separated graph components equals the multiplicity of the eigenvalue zero of L_{sym}. See, for example, [20] for the proof. Since our proteome graph is connected, our L_{sym} has the single eigenvalue zero.

- From (2.86), $\mathrm{rank}(A) = \mathrm{rank}(\hat{P})$. Here, it is obvious that $\mathrm{diag}(P^\top P)^{-1/2}$ has full rank, hence, $\mathrm{rank}(\hat{P}) = \mathrm{rank}(P) = m$. Therefore, $\mathrm{rank}(A) = m$. Obviously, D has full rank by definition and thus, $\mathrm{rank}(D^{-1/2}AD^{-1/2}) = \mathrm{rank}(A) = m$. Therefore, $D^{-1/2}AD^{-1/2}$ has $n-m$ singular values of zero. Since $D^{-1/2}AD^{-1/2}$ is symmetric, its singular values and eigenvalues are the same. Thus, $D^{-1/2}AD^{-1/2}$ has $n-m$ singular values of zero. Therefore, $L_{\mathrm{sym}}(= I - D^{-1/2}AD^{-1/2})$ has $n-m$ singular values of $1(= 1-0)$.

- For any n-dimensional vector $\boldsymbol{x}$, $\boldsymbol{x}^\top D^{-1/2}AD^{-1/2}\boldsymbol{x} = \left(\hat{P}D^{-1/2}\boldsymbol{x}\right)^\top \hat{P}D^{-1/2}\boldsymbol{x} = \left\|\hat{P}D^{-1/2}\boldsymbol{x}\right\|_2^2 \geq 0$. Therefore, $D^{-1/2}AD^{-1/2}$ is positive semi-definite, and all of the eigenvalues of $L_{\mathrm{sym}}(= I - D^{-1/2}AD^{-1/2})$ are less than or equal to one.

By these four points, we see that the eigenvalues of $L_{\mathrm{sym}}(= I - D^{-1/2}AD^{-1/2})$ satisfy

$$0 = \lambda_{\mathrm{LE}0} < \lambda_{\mathrm{LE}1} \leq \cdots \leq \lambda_{\mathrm{LE}(m-1)} < \lambda_{\mathrm{LE}m} = \lambda_{\mathrm{LE}(m+1)} = \cdots = \lambda_{\mathrm{LE}(n-1)} = 1. \tag{2.89}$$

Now we define an $m \times m$ matrix $\tilde{\Lambda}_{\mathrm{LE}}$ as

$$\tilde{\Lambda}_{\mathrm{LE}} = \begin{pmatrix} 0 & & & \\ & \lambda_{\mathrm{LE}1} & & \\ & & \ddots & \\ & & & \lambda_{\mathrm{LE}(m-1)} \end{pmatrix}. \tag{2.90}$$

Then, we can write

$$\Lambda_{\mathrm{LE}} = \begin{pmatrix} \tilde{\Lambda}_{\mathrm{LE}} & \\ & I \end{pmatrix}. \tag{2.91}$$

Let $\tilde{V}_{\mathrm{sym}}$ be the first m columns of V_{sym};

$$\tilde{V}_{\mathrm{sym}} = \begin{bmatrix} \hat{\boldsymbol{v}}_{\mathrm{sym},1} \\ \vdots \\ \hat{\boldsymbol{v}}_{\mathrm{sym},n} \end{bmatrix} = \begin{bmatrix} v_{\mathrm{sym},10} & \cdots & v_{\mathrm{sym},1(m-1)} \\ \vdots & \ddots & \vdots \\ v_{\mathrm{sym},n0} & \cdots & v_{\mathrm{sym},n(m-1)} \end{bmatrix}. \tag{2.92}$$

The truncated version of the eigenproblem is

$$L_{\mathrm{sym}}\tilde{V}_{\mathrm{sym}} = \tilde{V}_{\mathrm{sym}}\tilde{\Lambda}_{\mathrm{LE}}. \tag{2.93}$$

$\hat{\boldsymbol{v}}_{\mathrm{sym},i} := \begin{pmatrix} v_{\mathrm{sym},i0} & \cdots & v_{\mathrm{sym},i(m-1)} \end{pmatrix}$ is the i-th row of $\tilde{V}_{\mathrm{sym}}$ and provides a new m-dimensional representation of the i-th protein. This representation of proteome reflects distance be-

tween each protein pair in terms of cosine similarity. [3]

The eigenproblem (2.87) can be transformed to

$$D^{-1/2}AD^{-1/2}V_{\mathrm{sym}} = V_{\mathrm{sym}}M, \tag{2.97}$$

where

$$M := I - \Lambda_{\mathrm{LE}} = \begin{pmatrix} I - \tilde{\Lambda}_{\mathrm{LE}} & \\ & O \end{pmatrix}. \tag{2.98}$$

This means that the columns of V_{sym} are also the (normalized) eignenvectors of $D^{-1/2}AD^{-1/2}$, and M is the corresponding eigenvalue matrix. Defining an $m \times m$ matrix $\tilde{M}$ as

$$\tilde{M} = I - \tilde{\Lambda}_{\mathrm{LE}} = \begin{pmatrix} 1 & & & \\ & 1 - \lambda_{\mathrm{LE1}} & & \\ & & \ddots & \\ & & & 1 - \lambda_{\mathrm{LE}(m-1)} \end{pmatrix}, \tag{2.99}$$

we can write

$$M = \begin{pmatrix} \tilde{M} & \\ & O \end{pmatrix}. \tag{2.100}$$

[3] Let

$$\tilde{V}_{\mathrm{rw}} = D^{-1/2}\tilde{V}_{\mathrm{sym}}. \tag{2.94}$$

Then, this eigenproblem can also be regarded as the following generalized eigenproblem,

$$L\tilde{V}_{\mathrm{rw}} = D\tilde{V}_{\mathrm{rw}}\tilde{\Lambda}_{\mathrm{LE}}. \tag{2.95}$$

Remembering that $L_{\mathrm{rw}} = D^{-1}L$, we can further transform it into an eigneproblem

$$L_{\mathrm{rw}}\tilde{V}_{\mathrm{rw}} = \tilde{V}_{\mathrm{rw}}\tilde{\Lambda}_{\mathrm{LE}}. \tag{2.96}$$

This is the form of eigenproblem that we explained previously in 1.5. In this section, we discuss it later.

Note that

$$1 > 1 - \lambda_{\mathrm{LE1}} \geq \cdots \geq 1 - \lambda_{\mathrm{LE}(m-1)} > 0. \tag{2.101}$$

(2.97) is further transformed into

$$A = D^{1/2} V_{\mathrm{sym}} M \left(V_{\mathrm{sym}}\right)^{\top} D^{1/2}. \tag{2.102}$$

Comparing (2.102) and (2.85) leads to

$$\mathrm{diag}\left(P^{\top}P\right)^{-1/2} P^{\top}P\, \mathrm{diag}\left(P^{\top}P\right)^{-1/2} = D^{1/2} V_{\mathrm{sym}} M \left(V_{\mathrm{sym}}\right)^{\top} D^{1/2}. \tag{2.103}$$

Connecting Raman-proteome transformation coefficients and cosine similarity LE proteome

We consider $P^{\top}P$ in two ways. First, from (2.103),

$$\begin{aligned}
P^{\top}P &= \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} V_{\mathrm{sym}} M \left(V_{\mathrm{sym}}\right)^{\top} D^{1/2}\, \mathrm{diag}\left(P^{\top}P\right)^{1/2} \\
&= \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} V_{\mathrm{sym}} M^{1/2} \left(\mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} V_{\mathrm{sym}} M^{1/2}\right)^{\top} \\
&= \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \left(\begin{array}{c|c} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} & O \end{array}\right) \left(\mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \left(\begin{array}{c|c} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} & O \end{array}\right)\right)^{\top} \\
&= \left(\begin{array}{c|c} \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} & O \end{array}\right) \left(\begin{array}{c|c} \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} & O \end{array}\right)^{\top} \\
&= \left(\begin{array}{c|c} \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} & O \end{array}\right) \left(\begin{array}{c} \left(\mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2}\right)^{\top} \\ O \end{array}\right) \\
&= \mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2} \left(\mathrm{diag}\left(P^{\top}P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \tilde{M}^{1/2}\right)^{\top}.
\end{aligned} \tag{2.104}$$

Since

$$D^{-1/2}AD^{-1/2} = D^{-1/2}\hat{P}^\top\hat{P}D^{-1/2} \quad (\because (2.86)) \tag{2.105}$$

$$= \left(\hat{P}D^{-1/2}\right)^\top \hat{P}D^{-1/2}, \tag{2.106}$$

the diagonal elements of $\tilde{M}$, i.e., the positive eigenvalues of $D^{-1/2}AD^{-1/2}$, are the square of the singular values of $\hat{P}D^{-1/2}$. Compact SVD of $\hat{P}D^{-1/2}$ is expressed as

$$\hat{P}D^{-1/2} = U_{\mathrm{LE}}\Sigma_{\mathrm{LE}}\left(V_{\mathrm{LE}}\right)^\top, \tag{2.107}$$

where Σ_{LE} is an $m \times m$ diagonal matrix whose diagonal elements are the singular values in decreasing order from the upper left, and $\left(U_{\mathrm{LE}}\right)^\top U_{\mathrm{LE}} = \left(V_{\mathrm{LE}}\right)^\top V_{\mathrm{LE}} = I$. We then obtain

$$\Sigma_{\mathrm{LE}} = \tilde{M}^{1/2} = \begin{pmatrix} 1 & & & \\ & \sqrt{1-\lambda_{\mathrm{LE}1}} & & \\ & & \ddots & \\ & & & \sqrt{1-\lambda_{\mathrm{LE}(m-1)}} \end{pmatrix}. \tag{2.108}$$

Thus,

$$P^\top P = \mathrm{diag}\left(P^\top P\right)^{1/2} D^{1/2}\tilde{V}_{\mathrm{sym}}\Sigma_{\mathrm{LE}}\left(\mathrm{diag}\left(P^\top P\right)^{1/2} D^{1/2}\tilde{V}_{\mathrm{sym}}\Sigma_{\mathrm{LE}}\right)^\top. \tag{2.109}$$

On the other hand, from (2.44) and (2.46),

$$P^\top P = \left(R_E \left(B_E\right)^\top\right)^\top R_E \left(B_E\right)^\top = B_E \left(R_E\right)^\top R_E \left(B_E\right)^\top \tag{2.110}$$

$$= B_E \left(\Sigma_{R_E}\right)^2 B_E^\top \tag{2.111}$$

$$= B_E \Sigma_{R_E} \left(B_E \Sigma_{R_E}\right)^\top . \tag{2.112}$$

Therefore, comparing (2.109) and (2.112) yields

$$B_E \Sigma_{R_E} = \mathrm{diag}\left(P^\top P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \Sigma_{\mathrm{LE}} \Theta, \tag{2.113}$$

where Θ is an $m \times m$ orthogonal matrix. We define the estimate of B_E as

$$B_E^{\mathrm{est}} \coloneqq \mathrm{diag}\left(P^\top P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}}. \tag{2.114}$$

By this notation, (2.113) can be written as

$$B_E \Sigma_{R_E} = B_E^{\mathrm{est}} \Sigma_{\mathrm{LE}} \Theta. \tag{2.115}$$

Here, the left-hand side represents Raman-proteome linear transformation, whereas the right-hand side except for Θ is derived only from proteome data. Note that in order to derive (2.113), LDA does not need to be applied to Raman data because $P^\top P$ can be written in the form of $B_E \left(\Sigma_{R_E}\right)^2 \left(B_E\right)^\top$ even if LDA is not applied to Raman data.

Normalization with constant terms

Now we consider normalizing both sides of (2.115) by the first columns.

With (2.60) and (2.61), the left-hand side of (2.115) can be rewritten as

$$B_E \Sigma_{R_E} \quad = \quad \sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right) B_E^{\mathrm{norm}} \Sigma_{R_E}^{\mathrm{norm}}. \tag{2.116}$$

Likewise, for the right-hand side, the first column of B_E^{est} (the estimated constant term) is

$$\boldsymbol{b}_0^{\mathrm{est}} := B_E^{\mathrm{est}}\left[:,1\right] = B_E^{\mathrm{est}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathrm{diag}\left(P^\top P\right)^{1/2} D^{1/2} \tilde{V}_{\mathrm{sym}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{2.117}$$

The first column of $\tilde{V}_{\mathrm{sym}}$ is the normalized eigenvector corresponding to the eigenvalue zero of L_{sym}. [4]

By writing D as

$$D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}, \tag{2.118}$$

$$\tilde{V}_{\mathrm{sym}}\left[:,1\right] = \tilde{V}_{\mathrm{sym}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \left(\sum_{i=1}^{n} d_i\right)^{-1/2} D^{1/2} \mathbf{1}_n. \tag{2.119}$$

[4] By the definition of L, $L\mathbf{1}_n = D\mathbf{1}_n - A\mathbf{1}_n = \mathrm{diag}(D) - \mathrm{diag}(D) = \mathbf{0}_n$. Hence, in general, L has an eigenvalue zero and a corresponding eigenvector $\mathbf{1}_n$. Therefore, by the definition of L_{rw}, $L_{\mathrm{rw}}\mathbf{1}_n = D^{-1}L\mathbf{1}_n = \mathbf{0}_n$; L_{rw} also has an eigenvalue zero and a corresponding eigenvector $\mathbf{1}_n$. The eigenproblems $L_{\mathrm{rw}}\tilde{V}_{\mathrm{rw}} = \tilde{V}_{\mathrm{rw}}\tilde{\Lambda}_{\mathrm{LE}}$ and $L_{\mathrm{sym}}\tilde{V}_{\mathrm{sym}} = \tilde{V}_{\mathrm{sym}}\tilde{\Lambda}_{\mathrm{LE}}$ are equivalent because one can obtain the eigenproblem of L_{sym} by left multiplying both sides of the eigenproblem of L_{rw} by $D^{1/2}$ and that of L_{rw} by left multiplying both sides of the eigenproblem of L_{sym} by $D^{-1/2}$. Eigenvalues of L_{rw} and L_{sym} are identical, and $\tilde{V}_{\mathrm{rw}} = D^{-1/2}\tilde{V}_{\mathrm{sym}}$ holds for their eigenvectors. Therefore, L_{sym} has an eigenvalue zero and a corresponding eigenvector $D^{1/2}\mathbf{1}_n$. The multiplicity of the eigenvalue zero of our L_{sym} is one, and its corresponding eigenvector is limited to $D^{1/2}\mathbf{1}_n$.

Thus, (2.117) can be further transformed as

$$\begin{aligned}
\boldsymbol{b}_0^{\text{est}} = B_E^{\text{est}}[:,1] &= \left(\sum_{i=1}^{n} d_i\right)^{-1/2} \operatorname{diag}\left(P^\top P\right)^{1/2} D^{1/2} D^{1/2} \mathbf{1}_n \\
&= \left(\sum_{i=1}^{n} d_i\right)^{-1/2} \operatorname{diag}\left(P^\top P\right)^{1/2} D \mathbf{1}_n && (2.120) \\
&= \frac{1}{\left(\sum_{i=1}^{n} d_i\right)^{1/2}} \begin{pmatrix} d_1 \sqrt{(\boldsymbol{p}_1)^\top \boldsymbol{p}_1} \\ \vdots \\ d_n \sqrt{(\boldsymbol{p}_n)^\top \boldsymbol{p}_n} \end{pmatrix}. && (2.121)
\end{aligned}$$

Remembering that both $\operatorname{diag}\left(P^\top P\right)$ and D are diagonal, we obtain

$$\operatorname{diag}\left(\boldsymbol{b}_0^{\text{est}}\right) = \operatorname{diag}\left(B_E^{\text{est}}[:,1]\right) = \left(\sum_{i=1}^{n} d_i\right)^{-1/2} \operatorname{diag}\left(P^\top P\right)^{1/2} D. \tag{2.122}$$

Therefore, from (2.114) and (2.122), the "estimated coefficients" normalized with the "estimated constants" are

$$\begin{aligned}
B_E^{\text{est,norm}} &:= \operatorname{diag}\left(\boldsymbol{b}_0^{\text{est}}\right)^{-1} B_E^{\text{est}} && (2.123) \\
&= \operatorname{diag}\left(\boldsymbol{b}_0^{\text{est}}\right)^{-1} \operatorname{diag}\left(P^\top P\right)^{1/2} D^{1/2} \tilde{V}_{\text{sym}} && (2.124) \\
&= \left(\sum_{i=1}^{n} d_i\right)^{1/2} D^{-1} \operatorname{diag}\left(P^\top P\right)^{-1/2} \operatorname{diag}\left(P^\top P\right)^{1/2} D^{1/2} \tilde{V}_{\text{sym}} && (2.125) \\
&= \left(\sum_{i=1}^{n} d_i\right)^{1/2} D^{-1/2} \tilde{V}_{\text{sym}} && (2.126) \\
&= \left(\sum_{i=1}^{n} d_i\right)^{1/2} \tilde{V}_{\text{rw}}. && (2.127)
\end{aligned}$$

We remark that the eigenproblem of L_{rw}, i.e.,

$$L_{\mathrm{rw}}\tilde{V}_{\mathrm{rw}} = \tilde{V}_{\mathrm{rw}}\tilde{\Lambda}_{\mathrm{LE}} \tag{2.128}$$

is equivalent to solving a minimization problem

$$\arg\min \sum_{i,j} \|\hat{\boldsymbol{v}}_{\mathrm{rw},i} - \hat{\boldsymbol{v}}_{\mathrm{rw},j}\|_2^2 \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} \quad \text{subject to} \quad \left(\tilde{V}_{\mathrm{rw}}\right)^\top D\tilde{V}_{\mathrm{rw}} = I. \tag{2.129}$$

Thus, the closer $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are in terms of cosine similarity, the closer $\hat{\boldsymbol{v}}_{\mathrm{rw},i}$ and $\hat{\boldsymbol{v}}_{\mathrm{rw},j}$ (the i-th and j-th rows of $\tilde{V}_{\mathrm{rw}}$, respectively). [5] Remember that this property of $\hat{\boldsymbol{v}}_{\mathrm{rw},i}$ and $\hat{\boldsymbol{v}}_{\mathrm{rw},j}$ is analogous to that of $\hat{\boldsymbol{b}}_i^{\mathrm{norm}}$ and $\hat{\boldsymbol{b}}_j^{\mathrm{norm}}$ (the i-th and j-th rows of B_E^{norm}, respectively) as

[5] The relation between the minimization problem and the eigenproblem is the following. The objective function of the minimization problem is

$$\sum_{i,j} \|\hat{\boldsymbol{v}}_{\mathrm{rw},i} - \hat{\boldsymbol{v}}_{\mathrm{rw},j}\|_2^2 \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} = 2\sum_i d_i \|\hat{\boldsymbol{v}}_{\mathrm{rw},i}\|_2^2 - 2\sum_{i,j} A_{ij}\hat{\boldsymbol{v}}_{\mathrm{rw},i}\cdot\hat{\boldsymbol{v}}_{\mathrm{rw},j}. \tag{2.130}$$

Here,

$$\mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top D\tilde{V}_{\mathrm{rw}}\right) = \sum_i d_i \|\hat{\boldsymbol{v}}_{\mathrm{rw},i}\|_2^2, \tag{2.131}$$

$$\mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top A\tilde{V}_{\mathrm{rw}}\right) = \sum_{i,j} A_{ij}\hat{\boldsymbol{v}}_{\mathrm{rw},i}\cdot\hat{\boldsymbol{v}}_{\mathrm{rw},j}. \tag{2.132}$$

Therefore, (2.130) can be transformed into

$$\sum_{i,j} \|\hat{\boldsymbol{v}}_{\mathrm{rw},i} - \hat{\boldsymbol{v}}_{\mathrm{rw},j}\|_2^2 \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} = 2\,\mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top D\tilde{V}_{\mathrm{rw}}\right) - 2\,\mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top A\tilde{V}_{\mathrm{rw}}\right) \tag{2.133}$$

$$= 2\,\mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top L\tilde{V}_{\mathrm{rw}}\right). \tag{2.134}$$

Thus, the minimization problem is

$$\arg\min \quad \mathrm{tr}\left(\left(\tilde{V}_{\mathrm{rw}}\right)^\top L\tilde{V}_{\mathrm{rw}}\right) \quad \text{subject to} \quad \left(\tilde{V}_{\mathrm{rw}}\right)^\top D\tilde{V}_{\mathrm{rw}} = I. \tag{2.135}$$

This can be transformed into the generalized eigenproblem $L\tilde{V}_{\mathrm{rw}} = D\tilde{V}_{\mathrm{rw}}\tilde{\Lambda}_{\mathrm{LE}}$ by the method of Lagrange multipliers [33].

explained in 2.1.4 and 1.7.2.

By considering the first (upper-left) element of Σ_{LE} is one, the right side of (2.113) can be transformed into

$$
\begin{aligned}
B_E^{\mathrm{est}}\Sigma_{\mathrm{LE}}\Theta &= \mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right) B_E^{\mathrm{est,norm}}\Sigma_{\mathrm{LE}}\Theta && (2.136)\\
&= \left[\left(\sum_{i=1}^{n} d_i\right)^{-1/2} \mathrm{diag}\left(P^\top P\right)^{1/2} D\right]\left[\left(\sum_{i=1}^{n} d_i\right)^{1/2} \tilde{V}_{\mathrm{rw}}\right]\Sigma_{\mathrm{LE}}\Theta. && (2.137)
\end{aligned}
$$

Therefore, from (2.115), (2.116), (2.136), and (2.137),

$$
\begin{aligned}
&\sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right) B_E^{\mathrm{norm}}\Sigma_{R_E}^{\mathrm{norm}}\\
&\quad= \mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right) B_E^{\mathrm{est,norm}}\Sigma_{\mathrm{LE}}\Theta && (2.138)\\
&\quad= \left[\left(\sum_{i=1}^{n} d_i\right)^{-1/2} \mathrm{diag}\left(P^\top P\right)^{1/2} D\right]\left[\left(\sum_{i=1}^{n} d_i\right)^{1/2} \tilde{V}_{\mathrm{rw}}\right]\Sigma_{\mathrm{LE}}\Theta. && (2.139)
\end{aligned}
$$

This is the equation that links the normalized Raman-proteome coefficient proteome structure and the cosine similarity LE proteome structure.

Mathematical interpretation of the obtained equation

From (2.139), if the distributions of B_E^{norm} and $\tilde{V}_{\mathrm{rw}}$ are similar, the diagonal matrix Θ must be similar to the identity matrix because large off-diagonal elements of Θ makes lower dimensions "mix" much with the higher dimensions. In addition, the directions of diagonal matrices $\mathrm{diag}\left(\boldsymbol{b}_0\right)$ and $\mathrm{diag}\left(P^\top P\right)^{1/2} D$ must also be close to each other even if Θ is close to the identity matrix. Note that the first column of Θ also reflects the relation between $\boldsymbol{b}_0$ and $\boldsymbol{b}_0^{\mathrm{est}}$ $\left(\propto \mathrm{diag}\left(P^\top P\right)^{1/2} D\mathbf{1}_n\right)$.

The obtained relation between Raman-proteome normalized coefficient structure and cosine similarity LE structure is summarized in Table S10.

Note that the relation between $\Sigma_{R_E}^{\mathrm{norm}}$ and Σ_{LE} can change depending on normalization of V_{LDA}. However, the difference is not important for the spatial correspondence between the two structures because they only affect scaling of the axes. Rather, Λ'_{LDA} and $\tilde{M} = \left(\Sigma_{\mathrm{LE}}\right)^2$, which determine the order of columns of V_{LDA} and $\tilde{V}_{\mathrm{rw}}$, are important.

Table S10: **Mathematical relation between Raman-proteome coefficients and cosine similarity LE proteomes.**
The matrices in the left-hand side of (2.138) (a proteome structure based on Raman-proteome coefficients) and their counterparts in the right-hand side of (2.138) (a proteome structure obtained with cosine similarity LE) are listed.

Raman-omics coef. structure	**cos. sim. LE structure**	**size and type of matrix**	**Description**
B_E^{norm}	$\left(\sum_{i=1}^n d_i\right)^{1/2} \tilde{V}_{\mathrm{rw}}$ $\left(= B_E^{\mathrm{est,norm}}\right)$	$n \times m$ matrix	coefficients normalized by constants
I	Θ	$m \times m$ orthogonal matrix	orthogonal transformation
$m^{-1/2} \operatorname{diag}\left((\mathbf{1}_m)^\top P\right)$ $\left(= m^{1/2} \operatorname{diag}(\boldsymbol{b}_0)\right)$	$\left(\sum_{i=1}^n d_i\right)^{-1/2} \operatorname{diag}\left(P^\top P\right)^{1/2} D$ $\left(= \operatorname{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)\right)$	$n \times n$ diagonal matrix	constant terms
$\Sigma_{R_E}^{\mathrm{norm}}$	Σ_{LE}	$m \times m$ diagonal matrix	singular values

Application to main data

The normalized coefficient proteome structure in Figures 6A and 6C are the scatter plots between different columns of B_E^{norm}. The cosine similarity proteome structures in Figures 5L and 6D are the scatter plots between different columns of $\tilde{V}_{\mathrm{rw}}$. [6] On the basis of the results of the mathematical analysis, we compared B_E^{norm} and $B_E^{\mathrm{est,norm}} = \left(\sum_{i=1}^{n} d_i\right)^{1/2} \tilde{V}_{\mathrm{rw}}$ in Figure S9G.

The similarity between the projections of the two distinct omics structures onto low-dimensional subspaces suggests that Θ is close to the identity matrix ((2.138) and Table S10). Figures S9A and S9C–S9E show that the actual Θ is indeed significantly close to the identity matrix. This suggests that the major changes in cellular Raman spectra detectable by LDA reflect the major changes in the proteome characterized by LE based on stoichiometry balance (cosine similarity).

The structural similarity also suggests that directions of $\mathrm{diag}\left(\boldsymbol{b}_0\right)$ and $\mathrm{diag}\left(P^\top P\right)^{1/2} D$ are also similar ((2.138) and Table S10). In fact, Figure S9F confirmed good agreement between $m^{1/2}\,\mathrm{diag}\left(\boldsymbol{b}_0\right)$ and $\left(\sum_{i=1}^{n} d_i\right)^{-1/2} \mathrm{diag}\left(P^\top P\right)^{1/2} D = \mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)$. Since these two quantities are calculated only from proteome data, this agreement is a characteristics of proteome data. See 2.2 for further analyses and discussion on this point.

[6] In our data analysis, we calculated $\tilde{V}_{\mathrm{rw}}$ as $\tilde{V}_{\mathrm{rw}} = D^{-1/2}\tilde{V}_{\mathrm{sym}}$, where each column of $\tilde{V}_{\mathrm{sym}}$ was normalized.

2.2 Quantitative constraint on omics profiles

2.2.1 From agreement between $\sqrt{m}\boldsymbol{b}_0$ and $\boldsymbol{b}_0^{\mathrm{est}}$ to proportionality between L^1 norm/L^2 norm ratio and degree

We observed above that constant terms $\boldsymbol{b}_0$ and the estimated constant terms $\boldsymbol{b}_0^{\mathrm{est}}$ were strongly correlated (Figure S9F). It is of note that both $\boldsymbol{b}_0$ and $\boldsymbol{b}_0^{\mathrm{est}}$ can be calculated only from omics data. Specifically, from (2.69) and (2.121),

$$\boldsymbol{b}_0 \propto \begin{pmatrix} (\mathbf{1}_m)^\top \boldsymbol{p}_1 \\ \vdots \\ (\mathbf{1}_m)^\top \boldsymbol{p}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m p_{i1} \\ \vdots \\ \sum_{i=1}^m p_{in} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m |p_{i1}| \\ \vdots \\ \sum_{i=1}^m |p_{in}| \end{pmatrix} = \begin{pmatrix} \|\boldsymbol{p}_1\|_1 \\ \vdots \\ \|\boldsymbol{p}_n\|_1 \end{pmatrix}, \tag{2.140}$$

$$\boldsymbol{b}_0^{\mathrm{est}} \propto \begin{pmatrix} d_1\sqrt{(\boldsymbol{p}_1)^\top \boldsymbol{p}_1} \\ \vdots \\ d_n\sqrt{(\boldsymbol{p}_n)^\top \boldsymbol{p}_n} \end{pmatrix} = \begin{pmatrix} d_1\sqrt{\sum_{i=1}^m p_{i1}^2} \\ \vdots \\ d_n\sqrt{\sum_{i=1}^m p_{in}^2} \end{pmatrix} = \begin{pmatrix} d_1\|\boldsymbol{p}_1\|_2 \\ \vdots \\ d_n\|\boldsymbol{p}_n\|_2 \end{pmatrix}. \tag{2.141}$$

Here, $\|\boldsymbol{p}_j\|_1$ and $\|\boldsymbol{p}_j\|_2$ are the L^1 and L^2 norms of $\boldsymbol{p}_j$ and reflect only the expression property of protein j. On the other hand, the degree d_j is a measure for the relationships of protein j with the other proteins because d_j is the sum of cosine similarities, $d_j = \sum_{i=1}^n \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j}$.

The observed relation

$$\sqrt{m}\boldsymbol{b}_0 \approx \boldsymbol{b}_0^{\mathrm{est}} \tag{2.142}$$

(Figure S9F) indicates that a proportionality relation

$$\frac{\|\boldsymbol{p}_j\|_1}{\|\boldsymbol{p}_j\|_2} \propto \sum_{i=1}^n \cos\theta_{\boldsymbol{p}_i\boldsymbol{p}_j} \tag{2.143}$$

must hold approximately.

As mentioned in the main text, we refer the ratio of L^1 norm to L^2 norm in the left-hand side of (2.143) as expression generality score (g_j) because it can be interpreted as a measure of constancy and generality of the expression levels of the protein (See 1.9 and Figures S8A and S8B). When a protein is perfectly condition-specific and expressed only in a particular condition, its L^1 norm equals its L^2 norm, and the ratio takes the minimum value one. When a protein is expressed equally across all the conditions, its L^1 norm is greater than its L^2 norm and the ratio takes the maximum value, the square root of the number of conditions (Figures S8A and S8B). On the other hand, the right-hand side of (2.143), which we refer as stoichiometry conservation centrality (d_j) in the main text, measures to what extent protein j conserves its stoichiometry with the other proteins. Therefore, the proportionality relation (2.143) suggests a global quantitative constraint between condition-specificity of expression patterns and stoichiometry conservation strength. Positions of SCGs and density of genes in the cosine similarity LE structure (Figures 5A, 5K, and 5L) already suggested that genes with less condition-specific expression patterns have more genes with stoichiometrically similar expression patterns, and the proportionality here quantitatively captures this property of omics dynamics. We remark that the proportionality relation was confirmed in all the omics data we analyzed in this paper (See 3.2 and Figures S7I–S7N).

2.2.2 Mathematics behind proportionality between stoichiometry conservation centrality d_j and expression generality score g_j

The cosine similarity-based analyses involve normalization of expression vectors $\boldsymbol{p}_j$ by its L^2 norm $\|\boldsymbol{p}_j\|_2$ (2.77). Normalized expression vectors $\boldsymbol{p}_j/\|\boldsymbol{p}_j\|_2$ represent points on the first orthant division of a unit $(m-1)$-sphere in a m-dimensional space ($\sum_{i=1}^{m}(x_i)^2 =$

1 $(x_1, \ldots, x_m \geq 0)$). L^2 normalization allows us to compare expression patterns without considering expression magnitudes, and an expression pattern is represented by a position on the unit $(m-1)$-sphere. Therefore, our cosine similarity-based quantification of stoichiometric balance in omics data is equivalent to evaluating distances (measured with angle) between points on the $(m-1)$-sphere.

Since stoichiometry conservation centrality d_j is the sum of cosine similarities,

$$d_j = \sum_{i=1}^{n} \cos\theta_{\boldsymbol{p}_j \boldsymbol{p}_i} = \sum_{i=1}^{n} \frac{\boldsymbol{p}_j \cdot \boldsymbol{p}_i}{\|\boldsymbol{p}_j\|_2 \|\boldsymbol{p}_i\|_2} \tag{2.144}$$

$$= \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \cdot \left(\sum_{i=1}^{n} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|_2} \right). \tag{2.145}$$

Defining

$$\tilde{\boldsymbol{p}}_{\mathrm{tot}} := \sum_{i=1}^{n} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|_2}, \tag{2.146}$$

we obtain

$$\frac{1}{\sqrt{\sum_{i=1}^{n} d_i}} d_j = \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \cdot \frac{\tilde{\boldsymbol{p}}_{\mathrm{tot}}}{\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2}. \tag{2.147}$$

Note that $\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2 = \sqrt{\sum_{i=1}^{n} d_i}$. The last term $\tilde{\boldsymbol{p}}_{\mathrm{tot}}/\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2$ is the normalized vector of the sum of all the normalized expression vectors, which we refer to as "expression-pattern norm vector." Therefore, (2.147) means that stoichiometry conservation centrality $d_j = \sum_j \cos\theta_{\boldsymbol{p}_j \boldsymbol{p}_i}$ is proportional to the cosine of the angle between the expression-pattern norm vector $\tilde{\boldsymbol{p}}_{\mathrm{tot}}/\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2$ and protein j's expression pattern $\boldsymbol{p}_j/\|\boldsymbol{p}_j\|_2$. The more distant the expression pattern of a protein is from that specified by the expression-pattern norm vector, the smaller d_j is.

On the other hand, expression generality score g_j is

$$g_j = \frac{\|\boldsymbol{p}_j\|_1}{\|\boldsymbol{p}_j\|_2} = \left\| \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \right\|_1 \tag{2.148}$$

$$= \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \cdot \mathbf{1}_m. \tag{2.149}$$

Therefore,

$$\frac{1}{\sqrt{m}} \frac{\|\boldsymbol{p}_j\|_1}{\|\boldsymbol{p}_j\|_2} = \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \cdot \frac{\mathbf{1}_m}{\sqrt{m}}. \tag{2.150}$$

The last term $\mathbf{1}_m/\sqrt{m}$ is the normalized vector of $\mathbf{1}_m$, corresponding to the "center" of the first orthant division of the unit $(m-1)$-sphere. In other words, $\mathbf{1}_m/\sqrt{m}$ represents "perfectly-even expression pattern" across conditions. Therefore, (2.150) means that the expression generality score $g_j = \|\boldsymbol{p}_j\|_1/\|\boldsymbol{p}_j\|_2$ is proportional to the cosine of the angle between the "perfectly-even expression pattern" $\mathbf{1}_m/\sqrt{m}$ and protein j's expression pattern $\boldsymbol{p}_j/\|\boldsymbol{p}_j\|_2$. The more distant the expression pattern of a protein is from the "perfectly-even expression pattern", the smaller the expression generality score is.

Comparing (2.147) and (2.150), we see that if the expression-pattern norm vector $\tilde{\boldsymbol{p}}_{\text{tot}}/\|\tilde{\boldsymbol{p}}_{\text{tot}}\|_2$ and the perfectly-even expression pattern $\mathbf{1}_m/\sqrt{m}$ are equal, a proportional relationship

$$\forall j,\ d_j = \sqrt{\frac{\sum_{i=1}^n d_i}{m}} \frac{\|\boldsymbol{p}_j\|_1}{\|\boldsymbol{p}_j\|_2} \tag{2.151}$$

holds. Note that this is equivalent to $\sqrt{m}\boldsymbol{b}_0 = \boldsymbol{b}_0^{\text{est}}$.

Conversely, if $\tilde{\boldsymbol{p}}_{\text{tot}}/\|\tilde{\boldsymbol{p}}_{\text{tot}}\|_2$ deviates from $\mathbf{1}_m/\sqrt{m}$, the proportional relation between d_j and g_j breaks. We found that the proteome data by Schmidt *et al.*[1] showed $\tilde{\boldsymbol{p}}_{\text{tot}}/\|\tilde{\boldsymbol{p}}_{\text{tot}}\|_2 \neq \mathbf{1}_m/\sqrt{m}$. Instead, we found that the values of the elements of $\tilde{\boldsymbol{p}}_{\text{tot}}$ increased approxi-

mately linearly with the population growth rates under corresponding conditions (Figure S8D). Such a strong positive correlation between the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ and the population growth rates is nontrivial and suggests a new growth law constraining the total of relative expression level changes of all the proteins.

Next, we consider the consequence of this positive correlation between the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ and the population growth rates. Let us consider the proteins whose expression generality score $g_j = \|\boldsymbol{p}_j\|_1 / \|\boldsymbol{p}_j\|_2$ takes the minimum value one. Namely, the expression of these proteins are completely condition-specific (Figures S8A and S8B). For such proteins, only one component of the expression pattern vector $\boldsymbol{p}_j / \|\boldsymbol{p}_j\|_2$ is one, and the other components are zero. Thus, from (2.147), their stoichiometry conservation centrality d_j becomes proportional to the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ corresponding to the conditions under which they are expressed. Since the values of the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ are positively correlated with the population growth rates, d_j of completely condition-specific proteins also exhibits a positive correlation with the growth rates under the conditions accompanying their expression (Figure S8C). Such correlation can be confirmed for the proteins with nearly condition-specific expression patterns (PaaE, Asr, and DgoA in Figures 7B and 7C).

More generally, the deviation of d_j from the perfect proportionality line can be understood by the relation

$$\frac{d_j}{\sqrt{\sum_{i=1}^{n} d_i}} - \frac{\|\boldsymbol{p}_j\|_1}{\sqrt{m}\|\boldsymbol{p}_j\|_2} = \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \cdot \left(\frac{\tilde{\boldsymbol{p}}_{\mathrm{tot}}}{\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2} - \frac{\mathbf{1}_m}{\sqrt{m}} \right), \tag{2.152}$$

which can be derived from (2.147) and (2.150). Note that the values of the elements of the last term $\tilde{\boldsymbol{p}}_{\mathrm{tot}} / \|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2 - \mathbf{1}_m / \sqrt{m}$ also increase with the population growth rates, being positive under fast growth conditions and negative under slow growth conditions (Figure S8D). Therefore, when protein j tends to be expressed higher under fast growth conditions,

i.e., when the elements of $\boldsymbol{p}_j/\left\|\boldsymbol{p}_j\right\|_2$ corresponding to the fast growth conditions are relatively larger than those corresponding to the slow growth conditions, the left-hand side of (2.152) becomes positive, and its d_j resides above the perfect proportionality line. On the other hand, when protein j tends to be expressed higher under slow growth conditions, its d_j resides below the perfect proportionality line.

In the g_j-d_j plot in Figures 7A and 7B, we find several stretches of protein clusters above and below the perfect proportionality line. As expected from the argument above, each cluster corresponds to a group of proteins with similar expression patterns, and their positions relative to the proportionality line characterize the condition under which they are expressed the most (Figure 7C).

In summary, visualizing omics data by using the stoichiometry conservation centrality d_j and the expression generality score g_j allows us to systematically characterize the condition-dependent expression pattern of each protein on the basis of its position in the plot. Interestingly, this systematic characterization of gene expression patterns (the relation between d_j and g_j) was derived from our mathematical analyses of the correspondences between Raman and omics as we explained above.

3 Extended data analysis

3.1 Growth laws

3.1.1 Single-gene-level growth law

Bacterial growth law states that the total abundances of ribosomal components increase linearly with growth rate [34][35][36]. The homeostatic core (the largest SCG) identified in our analysis contains many ribosomal proteins. Hence, it is plausible that the total abundance of homeostatic core proteins also increases linearly with growth rate, which we indeed found true (Figure S5A). Furthermore, the abundance ratios of homeostatic core proteins are conserved across conditions. Therefore, the intracellular abundance of each protein species in the homeostatic core is expected to increase linearly with growth rate.

Let $p_{\epsilon j}$ be the abundance of protein j in the homeostatic core in environment ϵ. Since this protein conserves the stoichiometry with the other homeostatic core proteins across conditions,

$$\frac{p_{\epsilon i}}{p_{\epsilon j}} = \alpha_{ij} \tag{3.1}$$

in any environments ϵ (α_{ij} is the environment-independent abundance ratio of the homeostatic core protein i to protein j).

Let $M_\epsilon = \sum_i p_{\epsilon i} = p_{\epsilon j} \sum_i \alpha_{ij}$ be the total abundance of homeostatic core proteins in environment ϵ. The growth law for the homeostatic core is

$$M_\epsilon = a + b g_\epsilon, \tag{3.2}$$

where g_ϵ is the growth rate in environment ϵ; a is the y-intercept; and b is the slope of

the linear relation. Therefore,

$$p_{\epsilon j} = \frac{a}{\sum_i \alpha_{ij}} + \frac{b}{\sum_i \alpha_{ij}} g_\epsilon. \tag{3.3}$$

This shows that the abundance of homeostatic core proteins satisfies single-gene-level growth law.

3.1.2 Extended verification of stoichiometry conservation

When the abundance ratios between protein i and protein j are conserved,

$$\frac{p_{cj}}{p_{ci}} = \frac{p_{sj}}{p_{si}}, \tag{3.4}$$

where c and s specify the environments (s signifies the standard environment). Hence,

$$\frac{p_{cj}}{p_{sj}} = \frac{p_{ci}}{p_{si}} = \gamma_c, \tag{3.5}$$

where γ_c is the common abundance ratio of stoichiometry-conserving proteins with respect to the condition c. Note that γ_c is common among all the proteins in a stoichiometry-conserving group.

From (3.5),

$$\log p_{cj} = \log p_{sj} + \log \gamma_c. \tag{3.6}$$

Therefore, plotting $\log p_{cj}$ against $\log p_{sj}$ should find the stoichiometry conserving proteins aligned on a straight line with a slope of 1. We indeed find such plots for the homeostatic core proteins (Figures S5B and S5C). This result confirms their stoichiometry conservation from a different perspective.

3.1.3 Linear dependence of common abundance ratio on growth rate

Since the total amount of homeostatic core proteins increase linearly with growth rate (Figure S5A and (3.2)),

$$\sum_j p_{cj} - \sum_j p_{sj} = b\left(g_c - g_s\right). \tag{3.7}$$

Since $\sum_j p_{cj} = \gamma_c \sum_j p_{sj}$,

$$\left(\gamma_c - 1\right) \sum_j p_{sj} = b\left(g_c - g_s\right). \tag{3.8}$$

Hence,

$$\gamma_c = \left(1 - \frac{bg_s}{M_s}\right) + \frac{b}{M_s} g_c, \tag{3.9}$$

where $M_s = \sum_j p_{sj}$. Therefore, the common abundance ratio γ_c also increases linearly with growth rate.

Estimating $\Gamma_c \coloneqq \log_{10} \gamma_c$ as the y-intercepts of the regression lines with a slope of 1 (Figure S5B), we confirmed this linear dependence of common abundance ratio of homeostatic core proteins on growth rate (Figure S5D).

3.2 Generality of the results

3.2.1 Additional datasets with Raman data

Correspondence of three types of space

The correspondences among LDA Raman, Raman-omimcs normalized coefficient omics structure and cosine similarity LE omics structure were also observed in other datasets. The datasets include Raman (this paper) and proteome [1] data of *E. coli* with different genotypes (BW25113, MG1655, and NCM3722) cultured in the "LB" medium (Figures S11A–S11E) and in the "Glucose" medium (Figures S11F–S11J), and Raman and transcriptome data of *S. pombe* cultured under 10 different environmental conditions (Figures S11K–S11O) [4].

Comparison of matrices obtained by mathematical analyses

In addition to the comparison of three types of space, we also examined the matrices on the basis of results of the mathematical analyses (Table S10) using the aforementioned additional datasets; the closeness of B_E^{norm} and $B_E^{\mathrm{est,norm}}$ (Figures S12F for Raman-proteome of *E. coli* with different genotypes cultured in "LB", S12L for Raman-proteome of *E. coli* with different genotypes cultured in "Glucose", and S12R for Raman-transcriptome of *S. pombe* cultured under in 10 environment conditions) , the closeness of Θ to the identity matrix (Figures S12A–S12D for Raman-proteome of *E. coli* with different genotypes cultured in "LB", S12G–S12J for Raman-proteome of *E. coli* with different genotypes cultured in "Glucose", and S12M–S12P for Raman-transcriptome of *S. pombe* cultured under in 10 environment conditions), and the correspondence between $\sqrt{m}\,\mathrm{diag}\,(\boldsymbol{b}_0)$ and $\mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)$ (Figures S12E for Raman-proteome of *E. coli* with different genotypes cultured

in "LB", S12K for Raman-proteome of *E. coli* with different genotypes cultured in "Glucose", and S12Q for Raman-transcriptome of *S. pombe* cultured under in 10 environment conditions). We confirmed that the same results hold for these additional datasets. Note that the correspondence between $\sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right)$ and $\mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)$ does not involve Raman data. It is an intrinsic property of the omics data.

Proportionality between expression generality and stoichiometry conservation centrality

The correspondence between $\sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right)$ and $\mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)$ suggested the proportionality between expression generality score g_j and stoichiometry conservation centrality d_j in these omics data. We confirmed that the same results hold for these additional datasets (Figures S7I–N).

Correlation between $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ and growth rates

For the proteome data of *E. coli* with different genotypes (BW25113, MG1655, and NCM3722) cultured in "LB" and in "Glucose", growth rates were also reported [1]. We confirmed a positive correlation between the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ and growth rates for these datasets (Figure S8E). See also the deviation from the proportionality line in the g_j-d_j plot (Figures S7I and S7J).

Biological relevance of centrality of cosine similarity LE structure

We also confirmed centrality-essentiality correlation and centrality-evolutionary conservation correlation in the *S. pombe* transcriptome data (Figures S6B and S6E–G).

Degree distribution and its destruction by randomization

Degree (stoichiometry conservation centrality) distributions of cosine similarity LE structure of the additional data sets also showed a similar pattern as the main data, and randomization of the omics data breaks the strong correlation of expression patterns in the actual data (Figures S7C, S7D, and S7H).

3.2.2 Additional datasets without Raman data

Examining proteome structures with cosine similarity LE does not require Raman data. Therefore, we additionally analyzed publicly available proteome data of *M. tuberculosis* and *M. bovis* under the growth conditions with distinct oxygen levels [2], and the proteome data of *S. cerevisiae* under various environmental conditions [3].

We characterized cosine similarity LE structures of these datasets (Figure S13). Furthermore, we confirmed the proportionality between expression generality score g_j and stoichiometry conservation centrality d_j (Figures S7K–M). For the proteome data of *S. cerevisiae*, growth rates were also reported [3]. The *S. cerevisiae* cells were cultured in chemostat at the same dilution rate in any condition. In fact, we observed little variation of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ (Figure S8F), which leads to little deviation from the proportionality line (Figure S7M).

Degree (stoichiometry conservation centrality) distributions of cosine similarity LE structure of the additional data sets also showed a similar pattern as the main data, and randomization of the omics data breaks the strong correlation of expression patterns in the actual data (Figures S7E–G).

4 Supplementary figures

Figure S1: **Schematic illustration of the approach in this study. Related to Figure 1.**

Raman spectra and gene expression profiles are both high-dimensional vectors and can be represented as points in high-dimensional spaces. Coarse-graining Raman spectra by dimensional reduction finds condition-dependent differences in their global spectral patterns (see Fig. 2). The dimension-reduced spectra were linked to and used to predict condition-dependent global gene expression profiles (see Fig. 2), which implies that global changes in spectral patterns detect differences in cellular physiological states. The analysis of this linkage led us to discover a stoichiometry-conserving constraint on gene expression, which enabled us to represent gene expression profiles in a functionally-relevant low-dimensional space ((i); see also Fig. 3–5). Then, we find a nontrivial correspondence between these low-dimensional Raman and gene expression spaces ((ii); see also Fig. 6). This correspondence provides an omics-level interpretation of global Raman spectral patterns and a quantitative constraint between expression generality and stoichiometry conservation centrality ((ii); see also Fig. S9 and Fig. 7).

Figure S2: **Custom-built Raman microscope and analyses of *E. coli* Raman spectra. Related to Figure 2.**

(A) Schematic diagram of the Raman microscope used in this study.

(B) Representative Raman spectra from single *E. coli* cells. The fingerprint region of one spectrum is shown for each condition.

(C) Linear superposition of Raman shifts. Each LDA axis is linear superposition of Raman shifts. These figures show the coefficients for LDA1 (left) and LDA2 (right).

(D) Relationship between Raman LDA1 axis and growth rates. The horizontal axis represents Raman LDA1 axis. The vertical axis represents growth rates measured in [1]. Each point corresponds to the data for one condition. Pearson's correlation coefficient is 0.81 ± 0.09.

Figure S3: **Estimation of proteomes from Raman spectra. Related to Figure 2.**

Comparing the measured proteomes with those estimated from Raman spectra. The horizontal and vertical axes represent the estimated and measured proteomes, respectively. Proteins with negative estimated abundance are not shown in these figures. The conditions with the largest and the second largest numbers of proteins with negative estimated abundance were "stationary3days" (666 proteins) and "LB" (359 proteins). The conditions with the fewest and the second fewest negatively estimated proteins were "GlucosepH6" (0 proteins) and "Xylose" (7 proteins).

Figure S4: **Comparison of stoichiometry conservation among COG classes. Related to Figure 3.**

(A and B) Relations between protein abundance and constant terms of Raman-proteome coefficients. The horizontal axes are $\boldsymbol{b}_0$ (constant terms), and the vertical axes are $\hat{\boldsymbol{p}}_i^\top$ (protein abundance). Dashed lines are the least squares regression lines with intercept zero for ISP COG class members. The average of B_{-i}^{est} was used as an estimate of B here. In (A), only ISP COG class members are shown for three representative conditions: "Galactose", "Glucose", and "GlycerolAA". In (B), all proteins are shown for a representative condition, "GlycerolAA".

(C) Relations between protein abundance and growth rates of *E. coli* under 15 environmental conditions. We analyzed the absolute quantitative proteome data, growth rate data, and COG annotation reported by [1]. Lines represent different protein species. Error bars are standard errors. The top panel is for the Cellular Processes and Signaling COG class; the middle is for the Information Storage and Processing COG class; and the bottom is for the Metabolism COG class.

(D) Relations between protein abundance and growth rates of three *E. coli* strains (BW25113, MG1655, and NCM3722) under two culture conditions. We again analyzed the data by [1]. Lines represent different protein species. Error bars are standard errors.

(E and F) COG class-dependent expression pattern similarity of *E. coli* proteomes between conditions. The *E. coli* proteome data under the 15 different environmental conditions were analyzed. The similarity is evaluated by Pearson correlation coefficients of log expression levels in (E) and by cosine similarity in (F). We consider all the combinations of the 15 conditions. Thus, there are 105 data points for each COG class. The box-and-whisker plots summarize the distributions of the points. The lines inside the boxes denote

the medians. The top and bottom edges of the boxes denote the 25th percentiles and 75th percentiles, respectively. Note that (E) and (F) are evaluation of the same data used in Figure 3B in the main text with different similarity indices.

(G) COG class-dependent expression pattern similarity between different strains of *E. coli* (BW25113, MG1655, and NCM3722). The absolute quantitative proteome data and COG annotation were taken from [1]. The similarity was evaluated by cosine similarity. The data contain three strains. Thus, there are thee points for each COG class. The top panel is for the "Glucose" condition; and the bottom is for the "LB" condition.

(H–J) COG class-dependent expression pattern similarity in other organisms. (H) is for *M. tuberculosis* (data from [2]; six environmental conditions (time points)), (I) for *M. bovis* (data from [2]; six environmental conditions (time points)), and (J) for *S. cerevisiae* (data from [3]; 10 environmental conditions). The COG annotations were taken from the December 2014 release of 2003-2014 COGs [37] and the Release 3 of "Mycobrowser" [38] for (H) and (I) and from the Comprehensive Sake Yeast Genome Database (S288C strain) [39] for (J). The unit for protein abundance was fg/cell for (H) and (I) and fg in pg dry cell weight for (J).

Figure S5: **Single-gene level growth law in the homeostatic core. Related to Figure 4.**

(A) Relationship between population growth rates and total abundance of SCG 1 (homeostatic core) proteins. Here, we analyzed the *E. coli* proteome data [1], focusing on the 15 conditions for which we obtained Raman data. The dashed line is the least squares regression line.

(B) Scatterplots of log abundance of SCG 1 (homeostatic core) proteins. Here, the proteomes under three representative conditions, "LB", "Glucose", and "Galactose", are compared with that under the standard condition "Glycerol." Each colored line is the linear regression line with slope one for the points with the same color. The vertical line is $x = 0$.

(C) Relationship between population growth rate and coefficient of determination of linear regression in (B). The vertical line represents the growth rate under the standard condition ("Glycerol").

(D) Linear relationship between common abundance ratio and growth rates. The vertical axis represents 10^{Γ_c} , where Γ_c is the y-intercepts in (B) (see 3.1.2). The dashed line is the linear regression line. The horizontal line is $y = 1$, and the x coordinate of the vertical line is the growth rate under the standard condition ("Glycerol").

(E) The gene loci of the proteins belonging to the condition-specific SCGs on the chromosome (ASM75055v1.46 [40]). Yellow dots are nodes (genes), and gray lines are edges (high cosine similarity relationships). The edge in the map of SCG 5 cannot be seen because their gene loci are clustered in close proximity in the same operon.

Figure S6: **Functional relevance of stoichiometry conservation centrality. Related to Figure 5.**

(A) Relationship between gene essentiality and stoichiometry conservation centrality in *E. coli*. The proportion of essential genes is plotted for each stoichiometry conservation centrality rank range. In this plot, we calculated stoichiometry conservation centrality based on the *E. coli* proteome data [1] under the 15 conditions for which we obtained Raman data. The list of essential genes was downloaded from EcoCyc [23].

(B) Relationship between gene essentiality and stoichiometry conservation centrality in *S. pombe*. We calculated stoichiometry conservation centrality based on the *S. pombe* transcriptome data reported in [4]. Only coding genes are considered in this plot, though stoichiometry conservation centrality values were calculated using both coding and non-coding genes. Gene classification is based on PomBase [24]. Some bins do not reach 100% in sum because eleven coding genes in the *S. pombe* transcriptome data were not found in the current PomBase.

(C) Relationship between ratio of coding genes and stoichiometry conservation centrality in the *S. pombe* transcriptome data. The coding/non-coding assignment is based on PomBase [24].

(D) Correlation between stoichiometry conservation and evolutionary conservation. In this plot, we calculated stoichiometry conservation centrality based on the *E. coli* proteome data [1] under the 15 conditions for which we obtained Raman data. Colors represent the height of each bar. The distributions of stoichiometry conservation centrality were compared between the top 25% and the bottom 25% fractions in the number of orthologs rankings. The fraction with many orthologs tends to have higher stoichiometry conservation centrality (one-sided Brunner–Munzel test, $p = 7.84 \times 10^{-15}$). The distri-

butions of the number of orthologs were compared between the top 25% and the bottom 25% stoichiometry conservation centrality fractions. The high centrality fraction tends to have more orthologs (one-sided Brunner–Munzel test, $p = 1.46 \times 10^{-11}$). Ortholog data were taken from OrthoMCL-DB [31].

(E–G) Correlation between stoichiometry conservation and evolutionary conservation in *S. pombe*. We calculated stoichiometry conservation centrality based on the *S. pombe* transcriptome data reported in [4]. In (E), the result is shown by two-dimensional histogram. Colors represent the height of each bar. The distributions of the number of orthologs were compared between the top 25% and the bottom 25% stoichiometry conservation centrality fractions. The high centrality fraction tends to have more orthologs (one-sided Brunner–Munzel test, $p = 0.00548$). The direct comparison between the two fractions is shown in (F). The distributions of stoichiometry conservation centrality were compared between the top 25% and the bottom 25% fractions in the number of orthologs rankings. The fraction with many orthologs tends to have higher stoichiometry conservation centrality (one-sided Brunner–Munzel test, $p = 0.00270$). The direct comparison between the two fractions is shown in (G). Ortholog data were taken from OrthoMCL-DB [31].

(H) Applying PCA to L^2-normalized proteomes. PCA (with mean-centering) was applied to L^2-normalized proteome data $\begin{bmatrix} \boldsymbol{p}_1/\|\boldsymbol{p}_1\|_2 & \cdots & \boldsymbol{p}_n/\|\boldsymbol{p}_n\|_2 \end{bmatrix}$. Here, we analyzed the *E. coli* proteome data under the 15 conditions for which we obtained Raman data. The left is a projection onto a two-dimensional space; and the right is a projection onto a three-dimensional space. The axes for visualization were selected by considering similarity to the cosine similarity LE structure.

Figure S7: **Distributions and constraints with respect to stoichiometry conservation centrality (degree). Related to Figures 5 and 7.**

(A) Comparison of degree (stoichiometry conservation centrality) distributions between original (yellow) and randomized (blue) *E. coli* proteome data. We created randomized proteome data by shuffling the expression levels across the protein species within each condition. We used the *E. coli* proteome data [1] under the 15 conditions for which we obtained Raman data.

(B) Comparison of the g_j-d_j relationships between original (yellow) and randomized data (blue). The horizontal axis is expression generality score ($g_j = L^1$ norm/L^2 norm) and the vertical axis is stoichiometry conservation centrality (d_j: degree). Each dot represents a protein species. The dashed lines are $y = n$, $x = 1, \sqrt{m}$ ($n = 2058, m = 15$). The solid lines are $y = \sqrt{\sum_i d_i/m}\ x$.

(C–H) Degree (stoichiometry conservation centrality) distributions for additional datasets. Yellow histograms are for the original data, and blue histograms are for the randomized data. (C) for the proteomes of three *E. coli* strains (BW25113, MG1655, and NCM3722) in LB [1]; (D) for the proteomes of the three *E. coli* strains in M9 Glucose [1]; (E) for the proteomes of *M. tuberculosis* [2]; (F) for the proteomes of *M. bovis* [2]; (G) for the proteomes of *S. cerevisiae* [3]; and (H) for the transcriptomes of *S. pombe* [4].

(I–N) g_j-d_j relationships for additional datasets. Each gray dots represent a protein species. The proteins belonging to the homeostatic core in each dataset are shown in magenta; those belonging to condition-specific SCGs are indicated in different colors in each plot. See the caption of Figures S11 and S13 for the cosine similarity threshold to specify the homeostatic core and the condition-specific SCGs in each dataset. The dashed

lines are $y = n, x = 1, \sqrt{m}$. The solid lines through the origins are $y = \sqrt{\sum_{i=1}^{n} d_i/m}\ x$. (I) for the proteomes of the three *E. coli* strains in LB [1]; (J) for the proteomes of the three *E. coli* strains in M9 Glucose [1]; (K) for the proteomes of *M. tuberculosis* [2]; (L) for the proteomes of *M. bovis* [2]; (M) for the proteomes of *S. cerevisiae* [3]; and (N) for the transcriptomes of *S. pombe* [4].

Figure S8: **Properties of normalized expression vectors. Related to Figure 7.**

(A and B) Schematic explanation for the interpretation of the L^1 norm/L^2 norm ratio of expression vectors as an index of expression generality. (A) is a two-dimensional case, and (B) is a three-dimensional case. The inset in (A) schematically explains L^1 norm and L^2 norm of an expression vector. See section 1.9 for detail.

(C) Schematic explanation for deviations of points from the proportionality line in the g_j-d_j plots. Here, we consider four condition-specific protein species a, b, c, and d labeled in the descending order of growth rates under the conditions accompanying their expression. Note that their L^1 norm/L^2 norm ratios are all one on the horizontal axis. One can show that the degree (stoichiometry conservation centrality) d_j is proportional to the inner product of L^2-normalized expression vector $\boldsymbol{p}_j/\|\boldsymbol{p}_j\|$ and the expression norm vector $\tilde{\boldsymbol{p}}_{\mathrm{tot}}/\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2$ (see (2.147) in section 2.2.2). Since the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}/\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2$ increase approximately linearly with growth rates of the corresponding conditions (see (D)), the degrees (stoichiometry conservation centrality values) decrease from a to d in the order of growth rates.

(D–F) Correlation between elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}$ and population growth rates. The vertical axis represents the elements of $\tilde{\boldsymbol{p}}_{\mathrm{tot}}/\|\tilde{\boldsymbol{p}}_{\mathrm{tot}}\|_2$ and the horizontal axis represents the population growth rates. The dashed lines are $y = 1/\sqrt{m}$. (D) is the result from the analysis of the *E. coli* proteome data [1] under the 15 conditions for which we obtained Raman data ($m = 15$). (E) is the result from the analysis of the proteome data of three strains of *E. coli* (BW25113, MG1655, and NCM3722) under "LB" and "Glucose" conditions ($m = 6$) [1]. (F) is the result from the analysis of the proteome data of *S. cerevisiae* under 10 different conditions ($m = 10$) [3]. The cells were cultured in chemostat with the

same dilution rate. The numbers of analyzable protein species and the numbers of conditions were different between (D) and (E). Thus, the values of the vertical axes cannot be compared directly between them.

Figure S9: **Mathematical analyses of the main Raman-proteome data. Related to Figure 6.**

Proteomes of *E. coli* under 15 conditions [1] and corresponding Raman data we measured in this study were analyzed in this figure.

(A) Visual comparison of the unit matrix I, the orthogonal matrix Θ obtained from the data, and a random orthogonal matrix. Height of each bar indicates the value of each element. Colors represent the height of each bar. For clarifying the position of each element, a component form of matrix Θ is shown in the middle ($m = 15$). For Θ (middle) and a random orthogonal matrix (right), the original matrices are displayed in the upper row, and matrices whose elements are the absolute values of the corresponding elements of the original matrices are displayed in the lower row. (In this figure, $|\Theta|$ represents a matrix of which the (i, j) element is the absolute value of the (i, j) element of Θ.)

(B) Representation of matrices as scatterplots. See section 1.8 for detail.

(C) Comparison of the unit matrix I, the orthogonal matrix Θ obtained from the data, and random orthogonal matrices Q by Pearson correlation coefficients. Pearson correlation coefficient of the element-wise squared matrix of each matrix can be regarded as a measure of closeness to the identity matrix ($\circ$ represents element-wise multiplication). The probability of finding a random orthogonal matrix Q with Pearson correlation coefficient greater than the Pearson correlation coefficient of Θ was $< 1 \times 10^{-5}$ (No occurrence in 10^5 samplings). See section 1.8 for detail.

(D) Comparison of magnitudes of off-diagonal elements among the unit matrix I, the orthogonal matrix Θ obtained from the data, and random orthogonal matrices Q. The lattice on the top explains the numbering of k-diagonals ($-m < k < m$, $m = 15$). In the lattices on the bottom, black color indicates areas in which the elements are squared and

summed at the corresponding steps (i.e., areas represented by x in the graph). The sum of the squared values in each step is shown in the middle graph. Error bars of the random matrix line are standard errors of 100 samplings. See section 1.8 for detail.

(E) Comparison of magnitudes of elements of leading principal submatrices among the unit matrix I, the orthogonal matrix Θ obtained from the data, and random orthogonal matrices Q. In the lattices on the bottom, black color indicates an area in which elements are squared and summed at the corresponding step (i.e., an area represented by x in the graphs). The sum of the squared values in each area is shown in the top graph. The results shown in the top graph are converted into ratios to the identity matrix I and are shown in the middle graph. Error bars of the random matrix line are standard errors of 100 samplings. See section 1.8 for detail.

(F) Comparison of $\sqrt{m}\,\mathrm{diag}\,(\boldsymbol{b}_0)$ and $\mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right)$. x axis represents $\sqrt{m}\boldsymbol{b}_0$ and y axis represents $\boldsymbol{b}_0^{\mathrm{est}}$. The dashed line indicates $y = x$.

(G) Comparison between B_E^{norm} (left) and $B_E^{\mathrm{est,norm}}$ (right). Note that while B_E^{norm} figure (left) is the same as Figure 6C, the right figure shows $B_E^{\mathrm{est,norm}} = (\sum_{i=1}^{n} d_i)^{1/2}\tilde{V}_{\mathrm{rw}}$, where $\tilde{V}_{\mathrm{rw}}$ is shown in Figure 6D.

Figure S10: **Orthant correspondences between Raman spectra in LDA space and condition-specific proteins in Raman-proteome coefficient proteome space. Related to Figure 6.**

Using the main Raman and proteome data of *E. coli* under the 15 conditions, we examine the orthant correspondence between Raman spectra in the LDA space and condition-specific proteins in the Raman-proteome coefficient proteome space Ω_{B}. Here, we focus on two proteins PaaE and AcrR.

(A) Expression patterns of PaaE (left) and AcrR (right) across conditions. Error bars are standard errors. PaaE is expressed under the "LB" condition in a condition-specific manner, whereas AcrR is expressed at high levels not only under "LB" condition but also under several other conditions.

(B) Positions of PaaE and AcrR in the Raman-proteome coefficient-based proteome space Ω_{B}.

(C) Verification of orthant correspondence. We verified the orthant correspondence described by the relation (2.76). We multiplied both sides of (2.76) by $\left(\Sigma_{R_E}^{\mathrm{norm}}\right)^{-1}$, and the elements of the vectors of both sides were compared by scatterplots. The horizontal axes are related to the coordinates in the Raman LDA space; the vertical axes are related to the coordinates in the Raman-proteome coefficient proteome space. The dashed lines are $y = x$. The nearly perfect agreement of the elements confirms the orthant correspondence for the condition-specific protein PaaE (left). Deviations from the diagonal agreement line are found for AcrR (right).

Figure S11: **Stoichiometry-based omics structures and their correspondences to Raman-based omics structures for additional datasets. Related to Figures 4–6.**

This figure summarizes the results on omics structures characterized by stoichiometry conservation relations and their correspondences to those characterized by Raman-omics relations for additional datasets. (A–E) show the results from the analyses of the Raman and proteome data of three *E. coli* strains (BW25113, MG1655, and NCM3722) in LB; (F–J) from the analyses of the Raman and proteome data of the three *E. coli* strains in M9 Glucose; and (K–O) from the analyses of the Raman and transcriptome data of *S. pombe* under 10 conditions. We used the *E. coli* proteome data reported in [1] and the *S. pombe* transcriptome data reported in [4] in the analyses. (A), (F), and (K) show distributions of omics components in cosine similarity LE space. Stoichiometry conservation centrality of each component is indicated by color. (B), (G), and (L) show expression patterns of representative condition-specific omics components indicated in the previous figures of omics structures in the csLE spaces. Error bars are standard errors in (B) and (G), and maximum-minimum ranges (two replicates) in (L). (C), (H), and (M) show positions of averaged cellular Raman spectra under different conditions in the LDA spaces. (D), (I), and (N) show omics structures in the spaces specified by the Raman-omics coefficients with the homeostatic cores and condition-specific SCGs indicated by colored points. (E), (J), and (O) show the omics structures in the csLE omics spaces with the homeostatic cores and condition-specific SCGs indicated by colored points. Columns $\boldsymbol{v}_{\mathrm{rw},1}$ (the eigenvector corresponding to L_{rw}'s smallest eigenvalue except for zero) and $\boldsymbol{v}_{\mathrm{rw},2}$ (the eigenvector corresponding to L_{rw}'s second smallest eigenvalue except for zero) are shown. We used the cosine similarity thresholds of 0.99993 to specify SCGs both for the three *E. coli*

strains under LB data ((D) and (E)) and for the three *E. coli* strains under M9 Glucose data ((I) and (J)), and 0.9967 for the *S. pombe* transcriptome data ((N) and (O)).

Figure S12: **Analyses of the mathematical relation connecting two types of omics spaces. Related to Figure 6.**

This figure shows the analyses of mathematical relation that connects coordinates of omics components in the two types of omics spaces (see Figure 6E and Section 2) using additional datasets. (A–F) show the results from the analyses of the Raman and proteome data of three *E. coli* strains (BW25113, MG1655, and NCM3722) in LB; (G–L) from the analyses of the Raman and proteome data of the three *E. coli* strains in M9 Glucose; and (M–R) from the analyses of the Raman and transcriptome data of *S. pombe* under 10 conditions. We used the *E. coli* proteome data reported in [1] and the *S. pombe* transcriptome data reported in [4] in the analyses. See the caption of Figure S9 for the explanation of each panel. The SCGs in (F), (L), and (R) are the same as in Figure S11. The probability of finding a random orthogonal matrix Q with Pearson correlation coefficient greater than the Pearson correlation coefficient of Θ was 0.022 in (B), 0.013 in (H), and $< 1 \times 10^{-5}$ (No occurrence in 10^5 samplings) in (N).

Figure S13: **Stoichiometry-based proteome structures for additional datasets. Related to Figures 4–5.**

This figure shows proteome structures in the csLE proteome spaces for additional datasets. (A–C) show the results from the analyses of the proteome data of *M. tuberculosis* H37Rv under gradual changes in oxygen levels [2]; (D–F) shows the results from the analyses of the proteome data of *M. bovis* BCG under gradual changes in oxygen levels [2]; and (G–I) show the results from the analyses of the proteome data of *S. cerevisiae* under 10 conditions in chemostat with the same dilution rate [3]. (A), (D), and (G) show the proteome structures in the csLE spaces. The thresholds used to specify the SCGs were 0.99965 for (A), 0.9997 for (D), and 0.9989 for (G). (B), (E), and (H) show the same proteome structures as in the previous panels, but with stoichiometry conservation centrality of each protein species indicated by the color. (C), (F), and (I) show expression patterns of representative proteins indicated by the red circles in the previous panels. Error bars in (C) are standard errors.

Figure S14: **Dependence of low-dimensional correspondence between Raman spectra and proteomes on the number of conditions. Related to Figure 6.**

The dependence of the low-dimensional correspondence between Raman spectra and proteomes on the number of analyzed conditions was systematically investigated by evaluating the similarity of the orthogonal matrix Θ to the identity matrix for all subsampled condition sets. Proteomes of *E. coli* under 15 conditions [1] and corresponding Raman data we measured in this study were analyzed in this figure.

(A) The relationship between the number of conditions and the probability of obtaining higher level of low-dimensional correspondence than that of experimental data by chance. This probability is calculated as the probability of finding a random orthogonal matrix with Pearson correlation coefficient greater than the Pearson correlation coefficient of Θ by creating 10^4 random orthogonal matrices. See section 1.8 and Figure S9 for details of the evaluation method. Each green square corresponds to one subsample, and each short horizontal black line represents the median of all the $\binom{15}{x}$ combinations of conditions (i.e., $\binom{15}{x}$ green squares) for each subsample size x. The blue dashed line indicates the detection limit (i.e., one over the number of generated random orthogonal matrices). The non-subsampled case (i.e., the case with all 15 conditions) in this figure corresponds to Figure S9C.

(B) Visual comparison of Θ, B_E^{norm}, and $B_E^{\mathrm{est,norm}}$ for six representative subsamples indicated in (A). As in Figure S9A, Θ is visualized using $|\Theta|$, whose element is the absolute value of the corresponding element of Θ, and height of each bar in the figures of $|\Theta|$ indicates the value of each element of $|\Theta|$. Colors reflect the height of each bar. Spaces created with columns of B_E^{norm} and $B_E^{\mathrm{est,norm}}$ are Ω_{B} and Ω_{LE}, respectively. As Θ deviates

from the identity matrix from the cases α and β to the case of ϵ, the low-dimensional correspondence between Ω_{B} and Ω_{LE} collapses naturally. Since the case ζ is the non-subsampled case, the figure of $|\Theta|$ is the same as Figure S9A, and those of B_E^{norm} and $B_E^{\mathrm{est,norm}}$ are the same as Figure S9G. Note that the figure of Ω_{B} of the case ζ is also exactly the same as Figure 6C, and that of Ω_{LE} of the case ζ is equal to Figure 6D up to a factor of $\left(\sum_{i=1}^{n} d_i\right)^{1/2}$. The SCGs shown in this figure were defined in the analysis of the proteomes of all the 15 conditions (Figure 4C).

References

[1] A. Schmidt, *et al.*, *Nature Biotechnology* **34**, 104 (2016).

[2] O. T. Schubert, *et al.*, *Cell Host & Microbe* **18**, 96 (2015).

[3] P.-J. Lahtvee, *et al.*, *Cell Systems* **4**, 495 (2017).

[4] K. J. Kobayashi-Kirschvink, *et al.*, *Cell Systems* **7**, 104 (2018).

[5] K. A. Datsenko, B. L. Wanner, *Proceedings of the National Academy of Sciences* **97**, 6640 (2000).

[6] T. Baba, *et al.*, *Molecular systems biology* **2**, 2006 (2006).

[7] F. R. Blattner, *et al.*, *Science* **277**, 1453 (1997).

[8] E. Soupene, *et al.*, *Journal of bacteriology* **185**, 5611 (2003).

[9] P. S. Swain, *et al.*, *Nature Communications* **7**, 1 (2016).

[10] A. D. Edelstein, *et al.*, *Journal of biological methods* **1** (2014).

[11] F. Huang, *et al.*, *Nature methods* **10**, 653 (2013).

[12] A. Savitzky, M. J. Golay, *Analytical chemistry* **36**, 1627 (1964).

[13] Y.-S. Huang, T. Karashima, M. Yamamoto, T. Ogura, H.-o. Hamaguchi, *Journal of Raman Spectroscopy* **35**, 525 (2004).

[14] T. De Bie, N. Cristianini, R. Rosipal, *Handbook of geometric computing* (Springer, 2005), pp. 129–167.

[15] R. Goodacre, *et al.*, *Microbiology* **144**, 1157 (1998).

[16] U. Nieminen, *Social Science Research* **2**, 371 (1973).

[17] S. Segarra, A. Ribeiro, *IEEE Transactions on Signal Processing* **64**, 543 (2015).

[18] S. Brin, L. Page, *Computer networks and ISDN systems* **30**, 107 (1998).

[19] P. Bonacich, *Journal of mathematical sociology* **2**, 113 (1972).

[20] U. Von Luxburg, *Statistics and computing* **17**, 395 (2007).

[21] M. Belkin, P. Niyogi, *Nips* (2001), vol. 14, pp. 585–591.

[22] M. Belkin, P. Niyogi, *Neural Computation* **15**, 1373 (2003).

[23] I. M. Keseler, *et al.*, *Nucleic Acids Research* **45**, D543 (2017).

[24] M. A. Harris, *et al.*, *Genetics* **220**, iyab222 (2022).

[25] M. A. Harris, A. Lock, J. Bähler, S. G. Oliver, V. Wood, *Bioinformatics* **29**, 1671 (2013).

[26] J. Cao, *et al.*, *Science* **370**, eaba7721 (2020).

[27] J. M. Replogle, *et al.*, *Cell* **185**, 2559 (2022).

[28] T. Wang, *et al.*, *Science* **350**, 1096 (2015).

[29] E. Brunner, U. Munzel, *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **42**, 17 (2000).

[30] T. Ara, *brunnermunzel: (Permuted) Brunner-Munzel Test* (2022). R package version 2.0.

[31] F. Chen, A. J. Mackey, C. J. Stoeckert Jr, D. S. Roos, *Nucleic Acids Research* **34**, D363 (2006).

[32] F. Mezzadri, *arXiv preprint math-ph/0609050* (2006).

[33] B. Ghojogh, F. Karray, M. Crowley, *arXiv preprint arXiv:1903.11240* (2019).

[34] F. C. Neidhardt, B. Magasanik, *Biochimica et Biophysica Acta* **42**, 99 (1960).

[35] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, T. Hwa, *Science* **330**, 1099 (2010).

[36] H. Bremer, P. P. Dennis, *EcoSal Plus* **3** (2008).

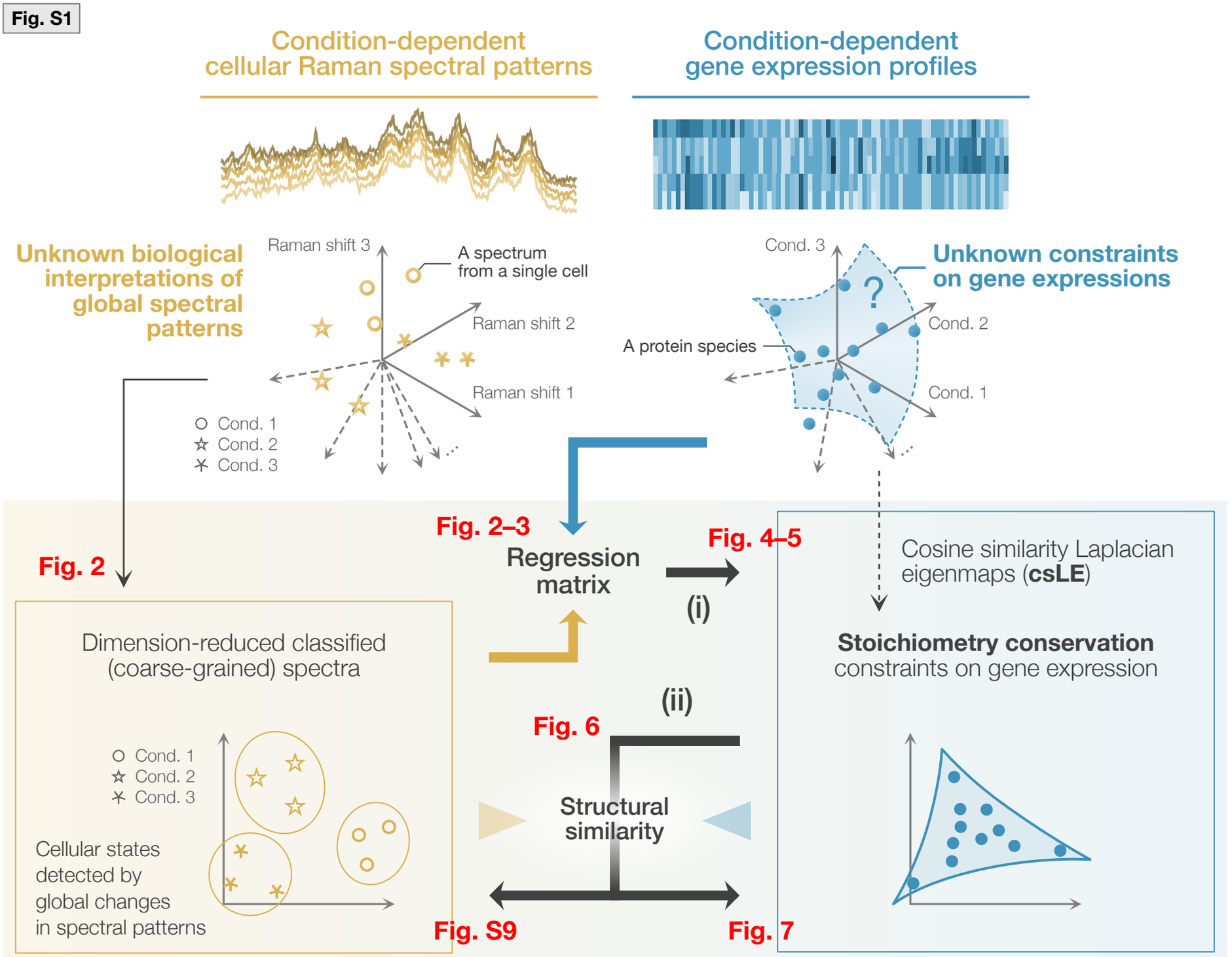
[37] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, *Nucleic Acids Research* **43**, D261 (2015).

[38] A. Kapopoulou, J. M. Lew, S. T. Cole, *Tuberculosis* **91**, 8 (2011).

[39] T. Akao, *et al.*, *DNA research* **18**, 423 (2011).

[40] K. L. Howe, *et al.*, *Nucleic Acids Research* **48**, D689 (2020).

Fig. S1



(i) develops csLE and uncovers stoichiometry conservation architecture of gene expressions, on which gene hierarchy correlates with its essentiality and evolutionary conservation.

(ii) finds a structural similarity, which indicates that spectra reflect the changes in omics profiles under the constraints of stoichiometry conservation and that stoichiometry conservation strength of each gene is proportional to expression generality score.

Fig. S2

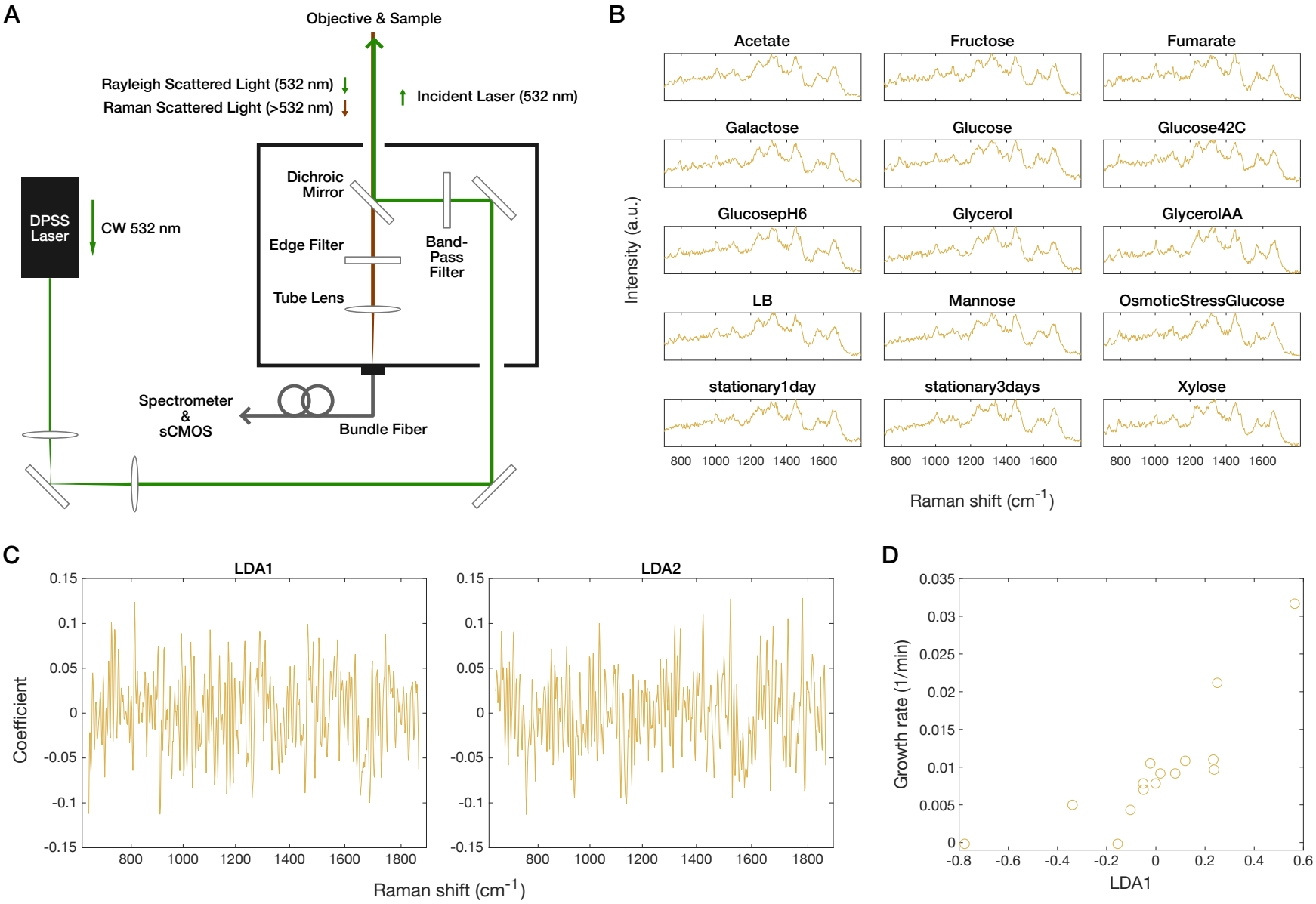

A
Objective & Sample
Rayleigh Scattered Light (532 nm)
Raman Scattered Light (>532 nm)
Incident Laser (532 nm)
DPSS Laser
CW 532 nm
Dichroic Mirror
Edge Filter
Band-Pass Filter
Tube Lens
Spectrometer & sCMOS
Bundle Fiber
B
Acetate
Fructose
Fumarate
Galactose
Glucose
Glucose42C
GlucosepH6
Glycerol
GlycerolAA
LB
Mannose
OsmoticStressGlucose
stationary1day
stationary3days
Xylose
Intensity (a.u.)
800 1000 1200 1400 1600
Raman shift (cm-1)
C
LDA1
LDA2
Coefficient
0.15 0.1 0.05 0 -0.05 -0.1 -0.15
800 1000 1200 1400 1600 1800
Raman shift (cm-1)
D
Growth rate (1/min)
0.035 0.03 0.025 0.02 0.015 0.01 0.005 0
-0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6
LDA1

Fig. S3

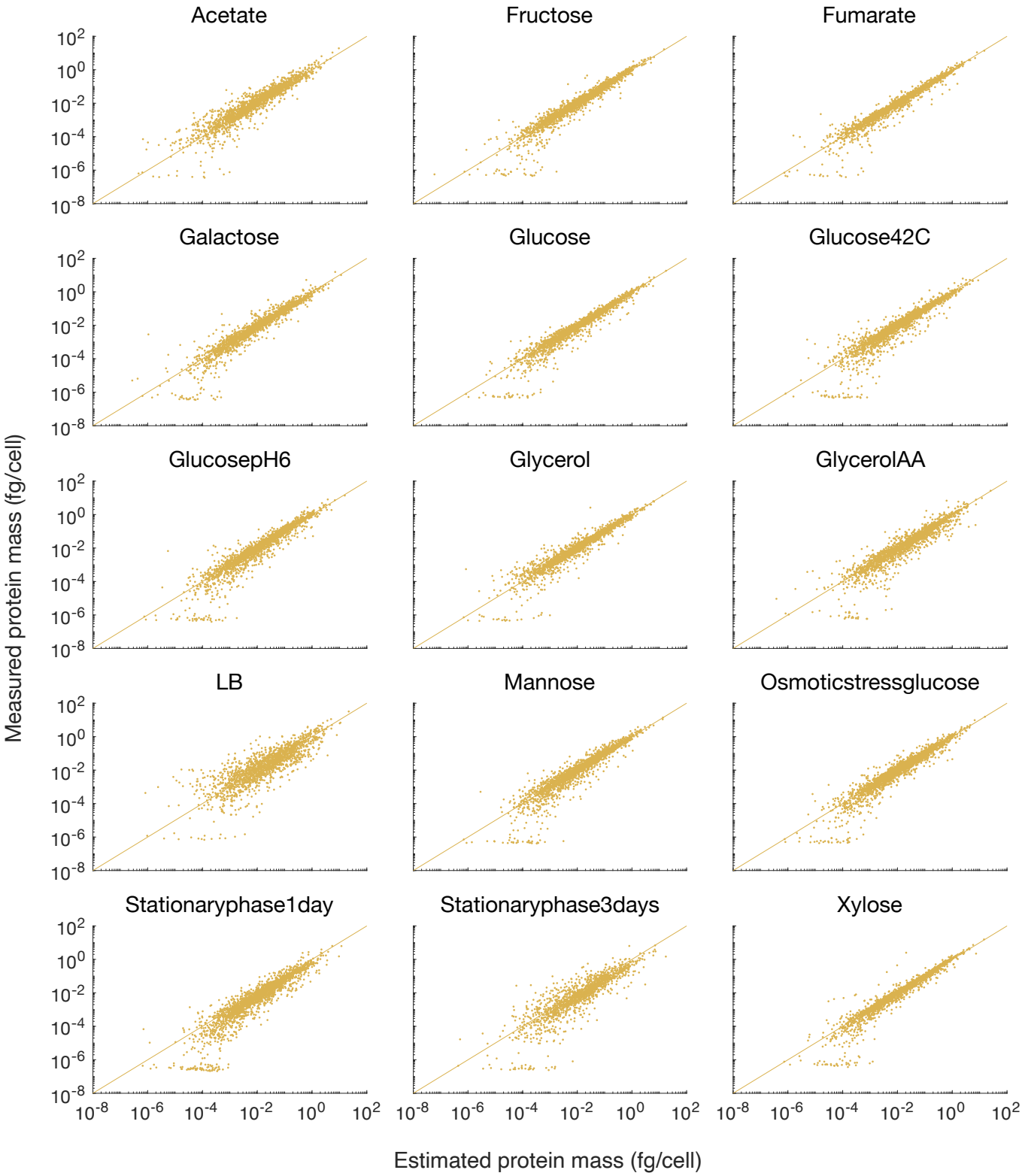

Acetate
Fructose
Fumarate
Galactose
Glucose
Glucose42C
GlucosepH6
Glycerol
GlycerolAA
LB
Mannose
Osmoticstressglucose
Stationaryphase1day
Stationaryphase3days
Xylose
Measured protein mass (fg/cell)
Estimated protein mass (fg/cell)
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}

Fig. S4

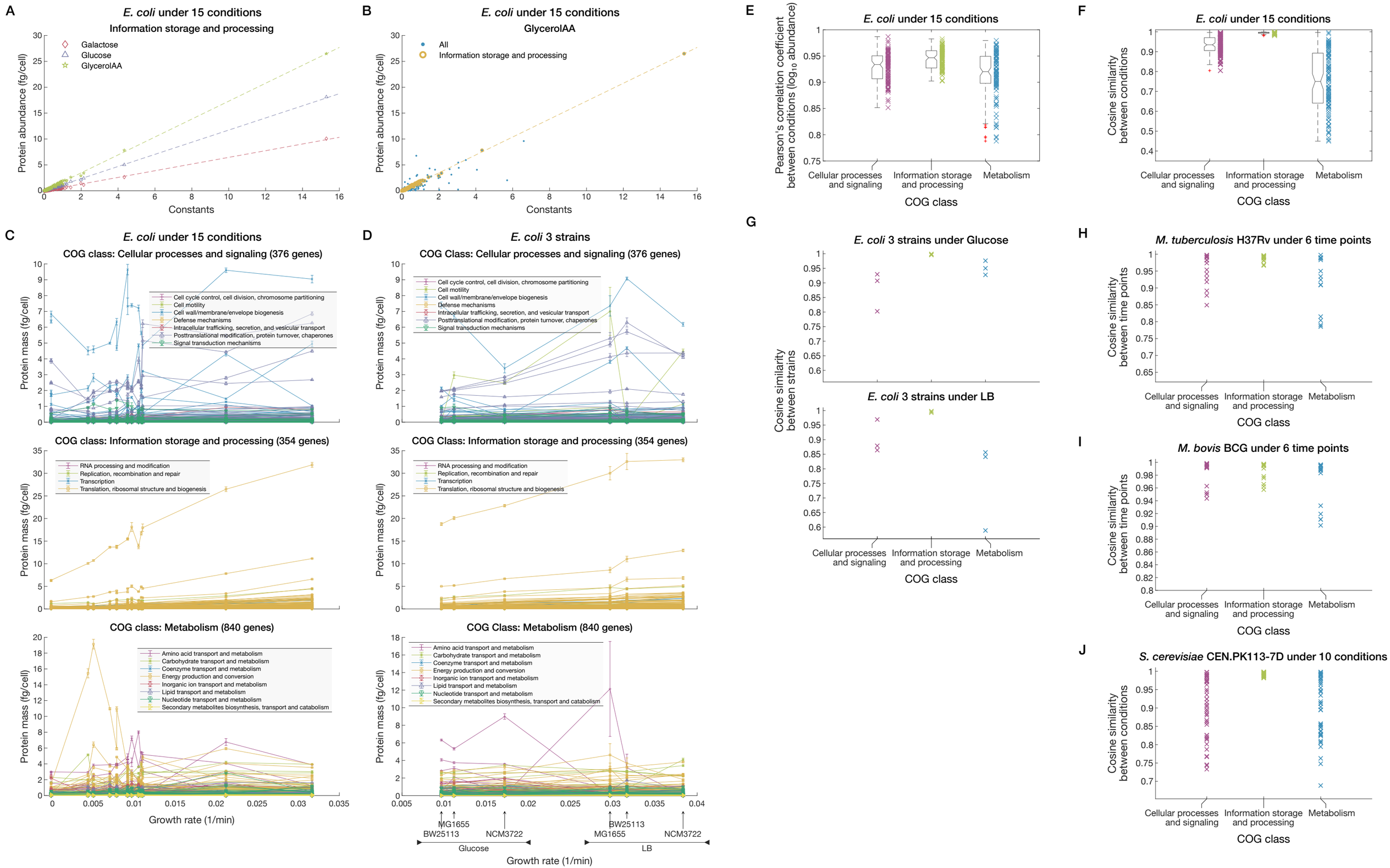

A
E. coli under 15 conditions
Information storage and processing
Galactose
Glucose
GlycerolAA
Protein abundance (fg/cell)
Constants
B
E. coli under 15 conditions
GlycerolAA
All
Information storage and processing
Protein abundance (fg/cell)
Constants
C
E. coli under 15 conditions
COG class: Cellular processes and signaling (376 genes)
Cell cycle control, cell division, chromosome partitioning
Cell motility
Cell wall/membrane/envelope biogenesis
Defense mechanisms
Intracellular trafficking, secretion, and vesicular transport
Posttranslational modification, protein turnover, chaperones
Signal transduction mechanisms
COG class: Information storage and processing (354 genes)
RNA processing and modification
Replication, recombination and repair
Transcription
Translation, ribosomal structure and biogenesis
COG class: Metabolism (840 genes)
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Coenzyme transport and metabolism
Energy production and conversion
Inorganic ion transport and metabolism
Lipid transport and metabolism
Nucleotide transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
Protein mass (fg/cell)
Growth rate (1/min)
D
E. coli 3 strains
COG Class: Cellular processes and signaling (376 genes)
COG Class: Information storage and processing (354 genes)
COG Class: Metabolism (840 genes)
MG1655
BW25113
NCM3722
Glucose
BW25113
MG1655
NCM3722
LB
Growth rate (1/min)
E
E. coli under 15 conditions
Pearson's correlation coefficient between conditions (log_{10} abundance)
Cellular processes and signaling
Information storage and processing
Metabolism
COG class
F
E. coli under 15 conditions
Cosine similarity between conditions
Cellular processes and signaling
Information storage and processing
Metabolism
COG class
G
E. coli 3 strains under Glucose
E. coli 3 strains under LB
Cosine similarity between strains
Cellular processes and signaling
Information storage and processing
Metabolism
COG class
H
M. tuberculosis H37Rv under 6 time points
Cosine similarity between time points
Cellular processes and signaling
Information storage and processing
Metabolism
COG class
I
M. bovis BCG under 6 time points
Cosine similarity between time points
Cellular processes and signaling
Information storage and processing
Metabolism
COG class
J
S. cerevisiae CEN.PK113-7D under 10 conditions
Cosine similarity between conditions
Cellular processes and signaling
Information storage and processing
Metabolism
COG class

Fig. S5

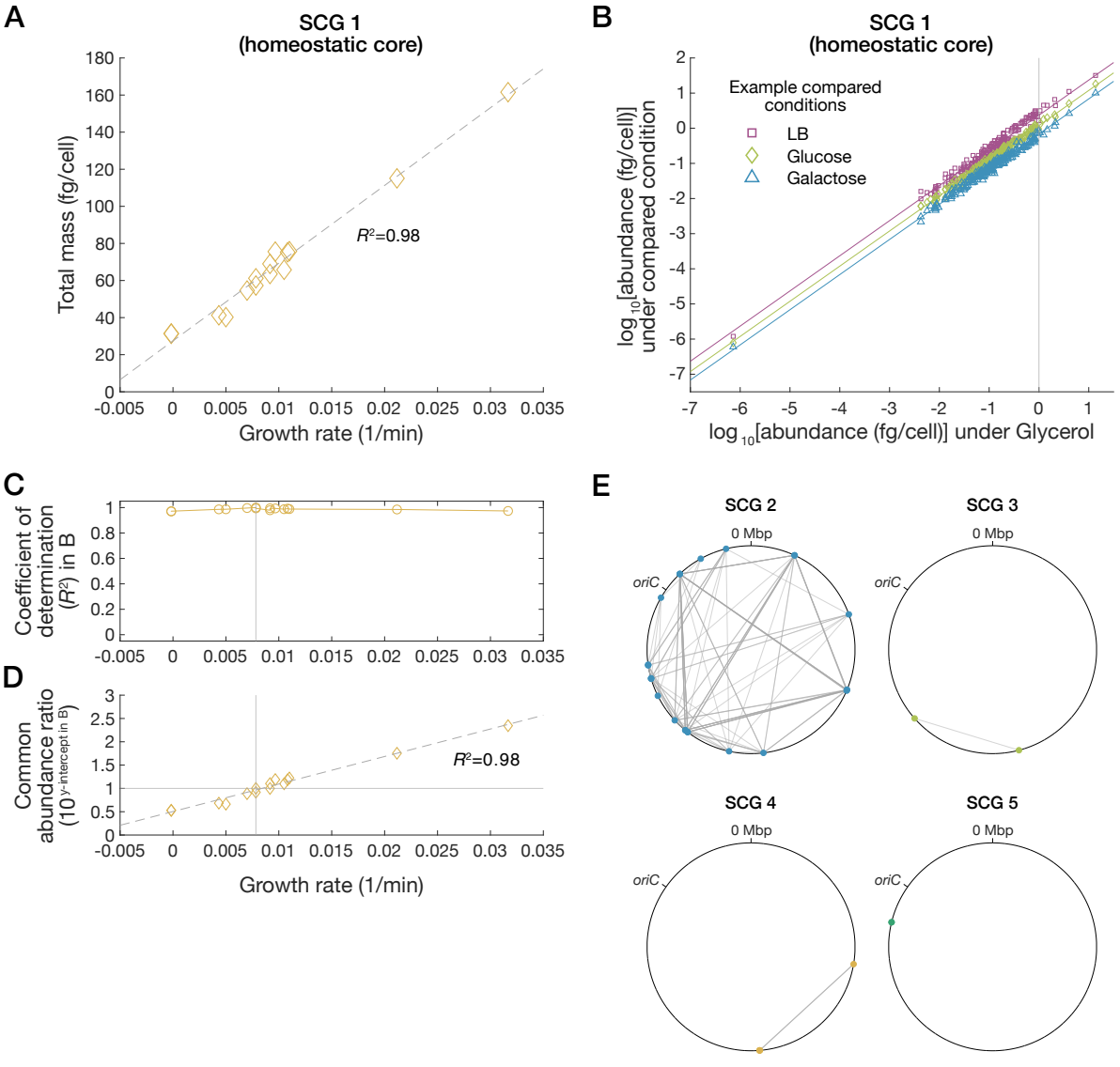

A
SCG 1
(homeostatic core)
Total mass (fg/cell)
Growth rate (1/min)
R²=0.98
B
SCG 1
(homeostatic core)
Example compared conditions
LB
Glucose
Galactose
log10[abundance (fg/cell)] under compared condition
log10[abundance (fg/cell)] under Glycerol
C
Coefficient of determination (R²) in B
D
Common abundance ratio (10^y-intercept in B)
R²=0.98
Growth rate (1/min)
E
SCG 2
SCG 3
SCG 4
SCG 5
0 Mbp
oriC

Fig. S6

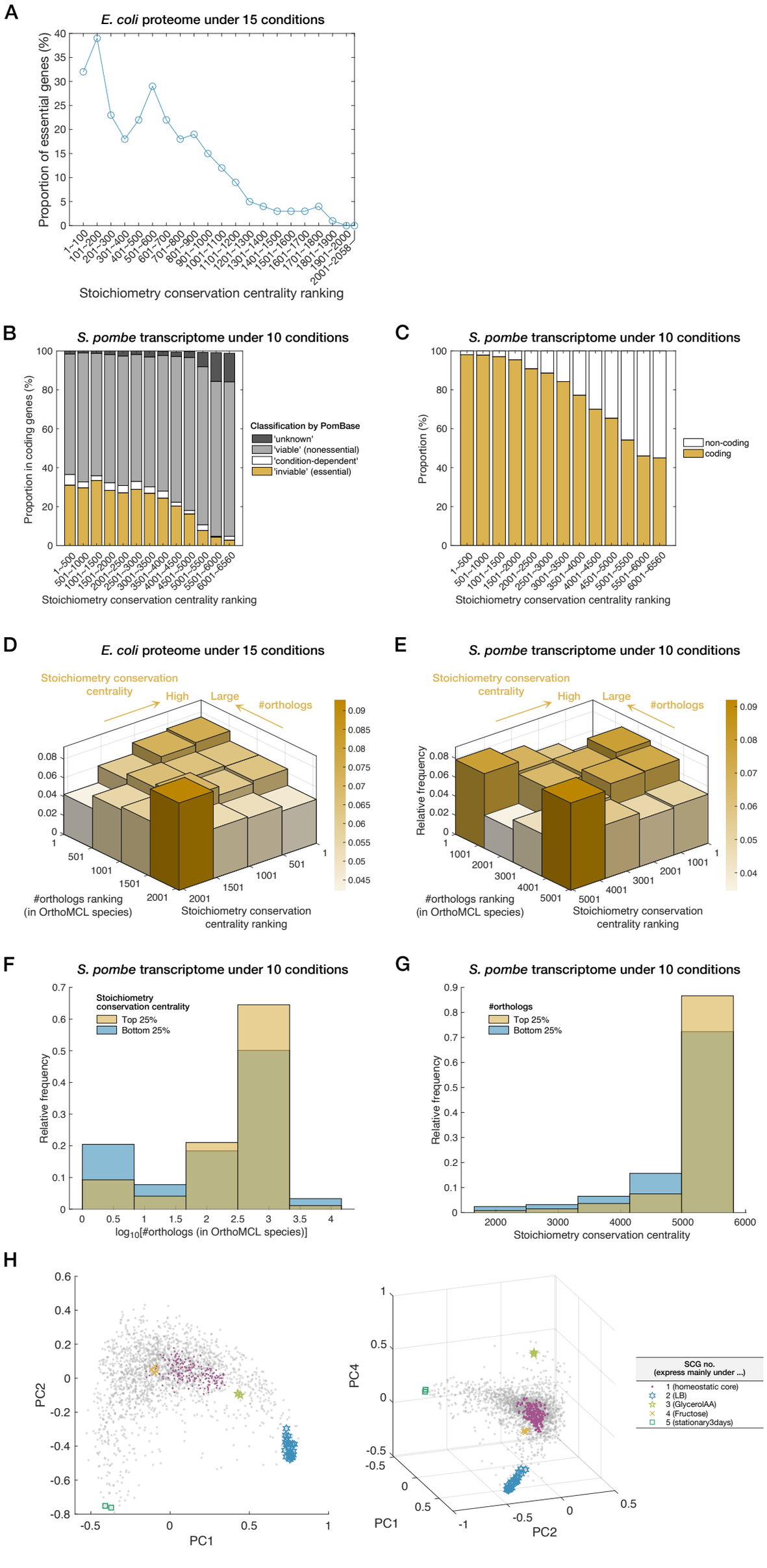

A
E. coli proteome under 15 conditions
Proportion of essential genes (%)
Stoichiometry conservation centrality ranking
B
S. pombe transcriptome under 10 conditions
Proportion in coding genes (%)
Classification by PomBase
'unknown'
'viable' (nonessential)
'condition-dependent'
'inviable' (essential)
Stoichiometry conservation centrality ranking
C
S. pombe transcriptome under 10 conditions
Proportion (%)
non-coding
coding
Stoichiometry conservation centrality ranking
D
E. coli proteome under 15 conditions
Stoichiometry conservation centrality
High
Large
#orthologs
#orthologs ranking (in OrthoMCL species)
Stoichiometry conservation centrality ranking
E
S. pombe transcriptome under 10 conditions
Relative frequency
#orthologs ranking (in OrthoMCL species)
Stoichiometry conservation centrality ranking
F
S. pombe transcriptome under 10 conditions
Stoichiometry conservation centrality
Top 25%
Bottom 25%
Relative frequency
log10[#orthologs (in OrthoMCL species)]
G
S. pombe transcriptome under 10 conditions
#orthologs
Top 25%
Bottom 25%
Relative frequency
Stoichiometry conservation centrality
H
PC1
PC2
PC4
SCG no.
(express mainly under ...)
1 (homeostatic core)
2 (LB)
3 (GlycerolAA)
4 (Fructose)
5 (stationary3days)

Fig. S7

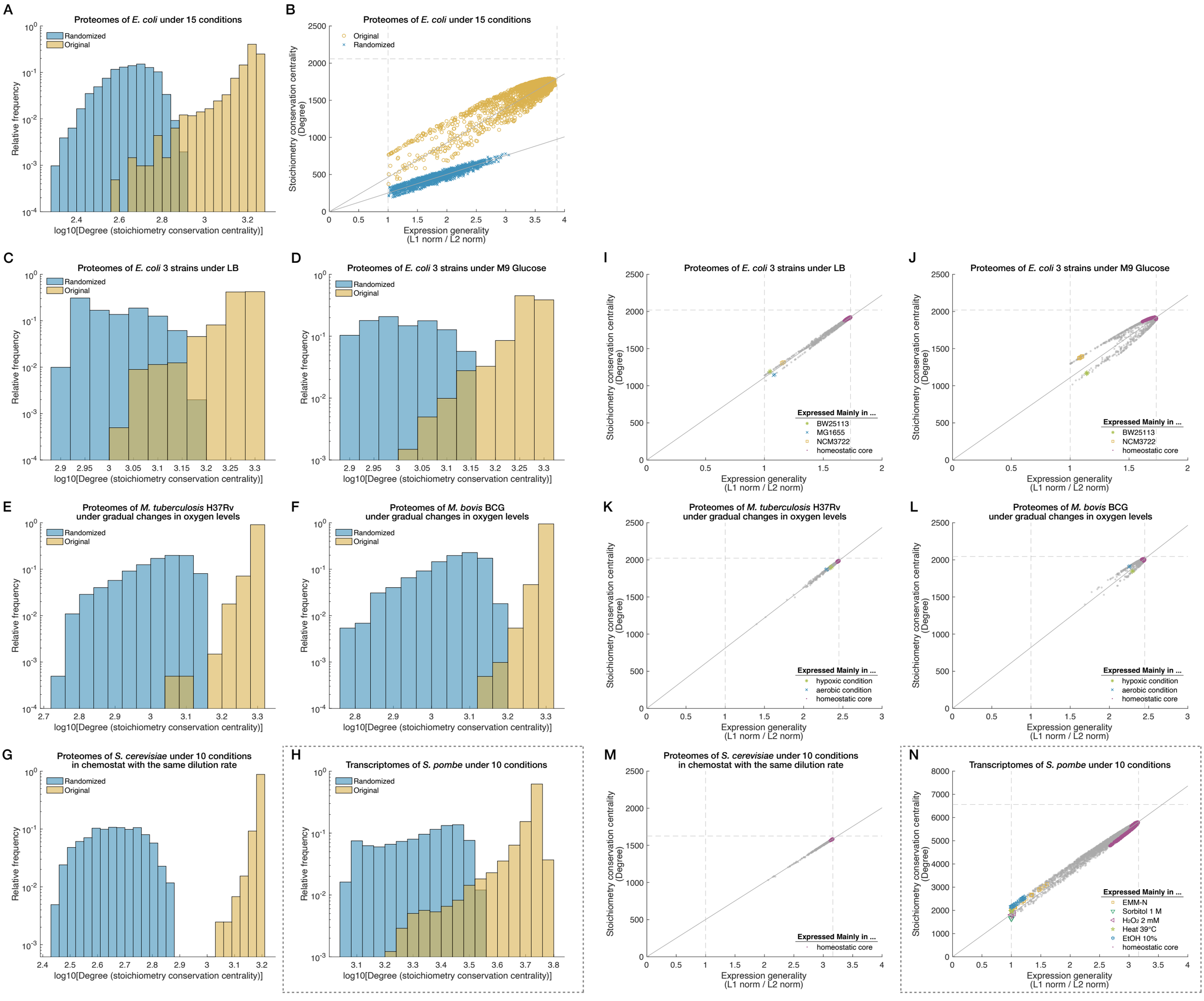

A
Proteomes of E. coli under 15 conditions
Randomized
Original
Relative frequency
log10[Degree (stoichiometry conservation centrality)]
B
Proteomes of E. coli under 15 conditions
Original
Randomized
Stoichiometry conservation centrality (Degree)
Expression generality (L1 norm / L2 norm)
C
Proteomes of E. coli 3 strains under LB
D
Proteomes of E. coli 3 strains under M9 Glucose
I
Proteomes of E. coli 3 strains under LB
J
Proteomes of E. coli 3 strains under M9 Glucose
Expressed Mainly in ...
BW25113
MG1655
NCM3722
homeostatic core
E
Proteomes of M. tuberculosis H37Rv under gradual changes in oxygen levels
F
Proteomes of M. bovis BCG under gradual changes in oxygen levels
K
L
hypoxic condition
aerobic condition
G
Proteomes of S. cerevisiae under 10 conditions in chemostat with the same dilution rate
H
Transcriptomes of S. pombe under 10 conditions
M
N
EMM-N
Sorbitol 1 M
H2O2 2 mM
Heat 39°C
EtOH 10%

Fig. S8

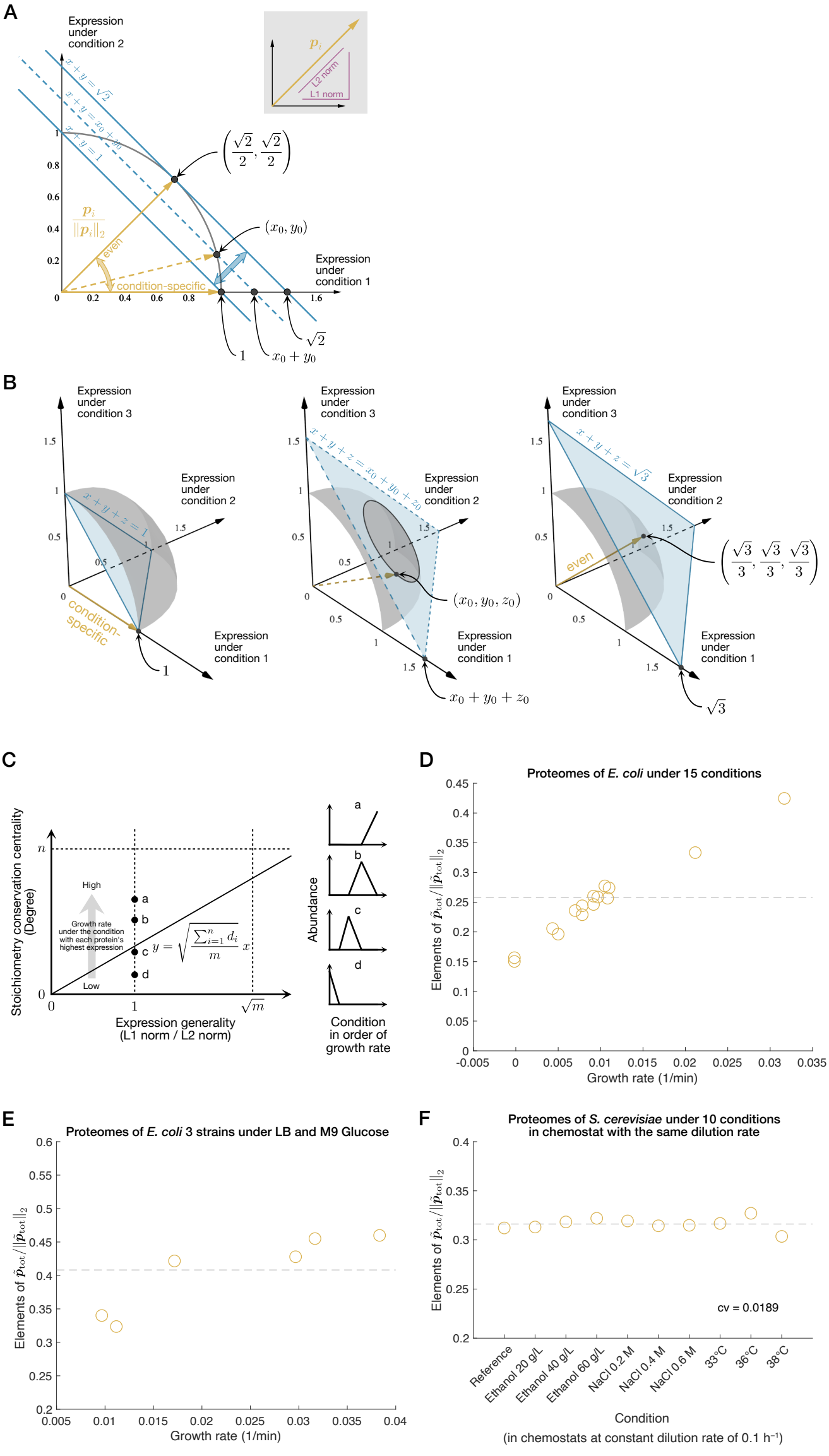
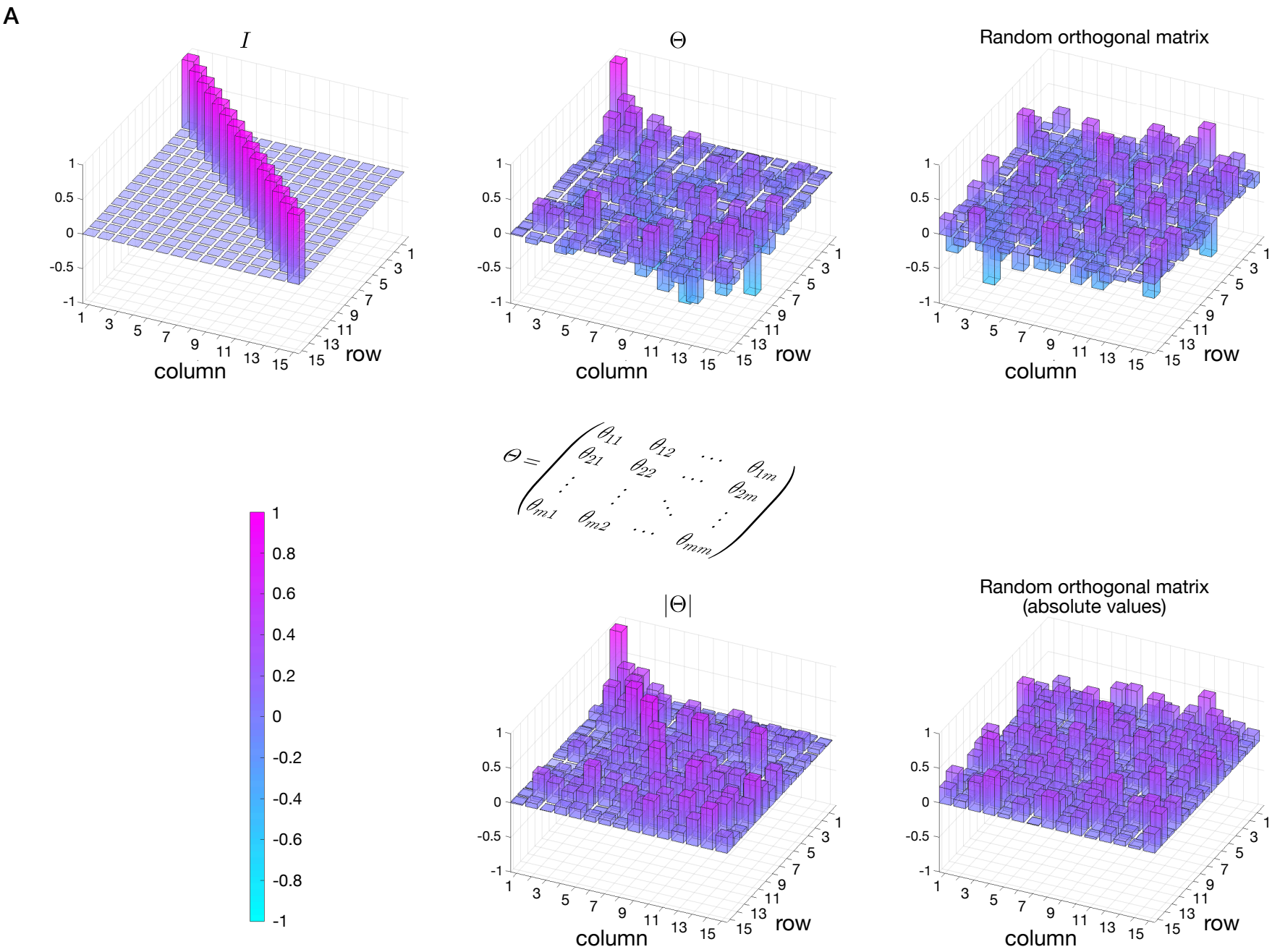
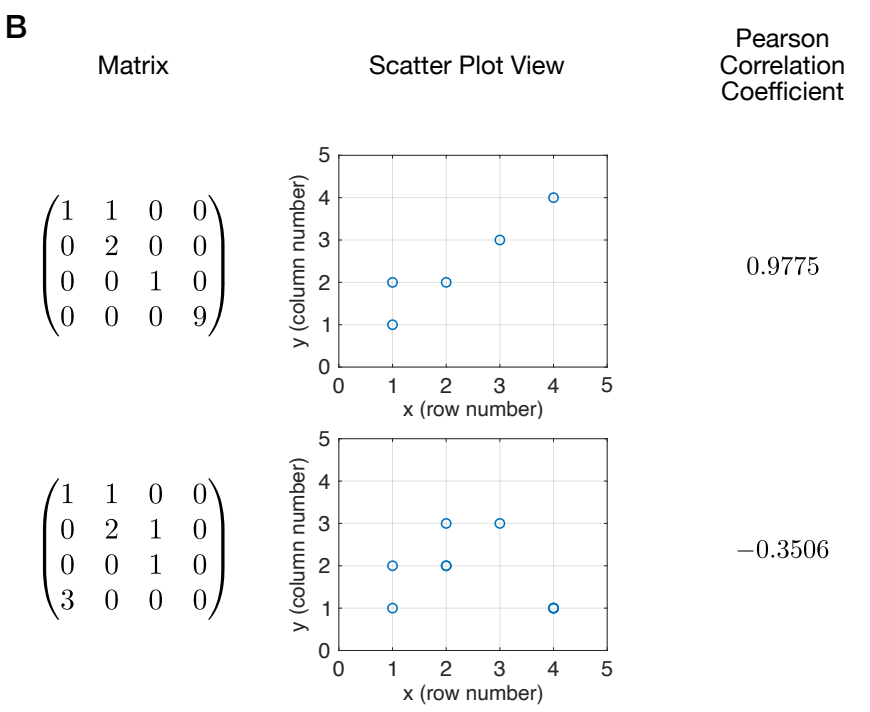

A
Expression under condition 2
Expression under condition 1
$x+y=\sqrt{2}$
$x+y=x_0+y_0$
$x+y=1$
$\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
(x_0,y_0)
$\frac{p_i}{\|p_i\|_2}$
even
condition-specific
p_i
L2 norm
L1 norm
1
x_0+y_0
$\sqrt{2}$
B
Expression under condition 3
Expression under condition 2
Expression under condition 1
$x+y+z=1$
condition-specific
$x+y+z=x_0+y_0+z_0$
(x_0,y_0,z_0)
$x_0+y_0+z_0$
$x+y+z=\sqrt{3}$
even
$\left(\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right)$
$\sqrt{3}$
C
Stoichiometry conservation centrality (Degree)
Expression generality (L1 norm / L2 norm)
High
Growth rate under the condition with each protein's highest expression
Low
$y=\sqrt{\frac{\sum_{i=1}^{n} d_i}{m}}\,x$
Abundance
Condition in order of growth rate
D
Proteomes of *E. coli* under 15 conditions
Elements of $\tilde{p}_{tot}/\|\tilde{p}_{tot}\|_2$
Growth rate (1/min)
E
Proteomes of *E. coli* 3 strains under LB and M9 Glucose
Elements of $\tilde{p}_{tot}/\|\tilde{p}_{tot}\|_2$
Growth rate (1/min)
F
Proteomes of *S. cerevisiae* under 10 conditions in chemostat with the same dilution rate
Elements of $\tilde{p}_{tot}/\|\tilde{p}_{tot}\|_2$
cv = 0.0189
Reference
Ethanol 20 g/L
Ethanol 40 g/L
Ethanol 60 g/L
NaCl 0.2 M
NaCl 0.4 M
NaCl 0.6 M
33°C
36°C
38°C
Condition
(in chemostats at constant dilution rate of 0.1 h⁻¹)

Fig. S9

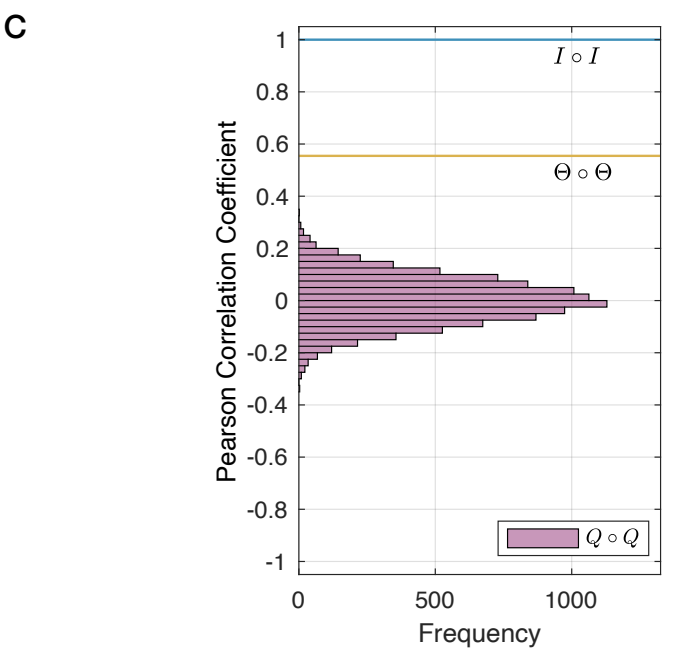
A



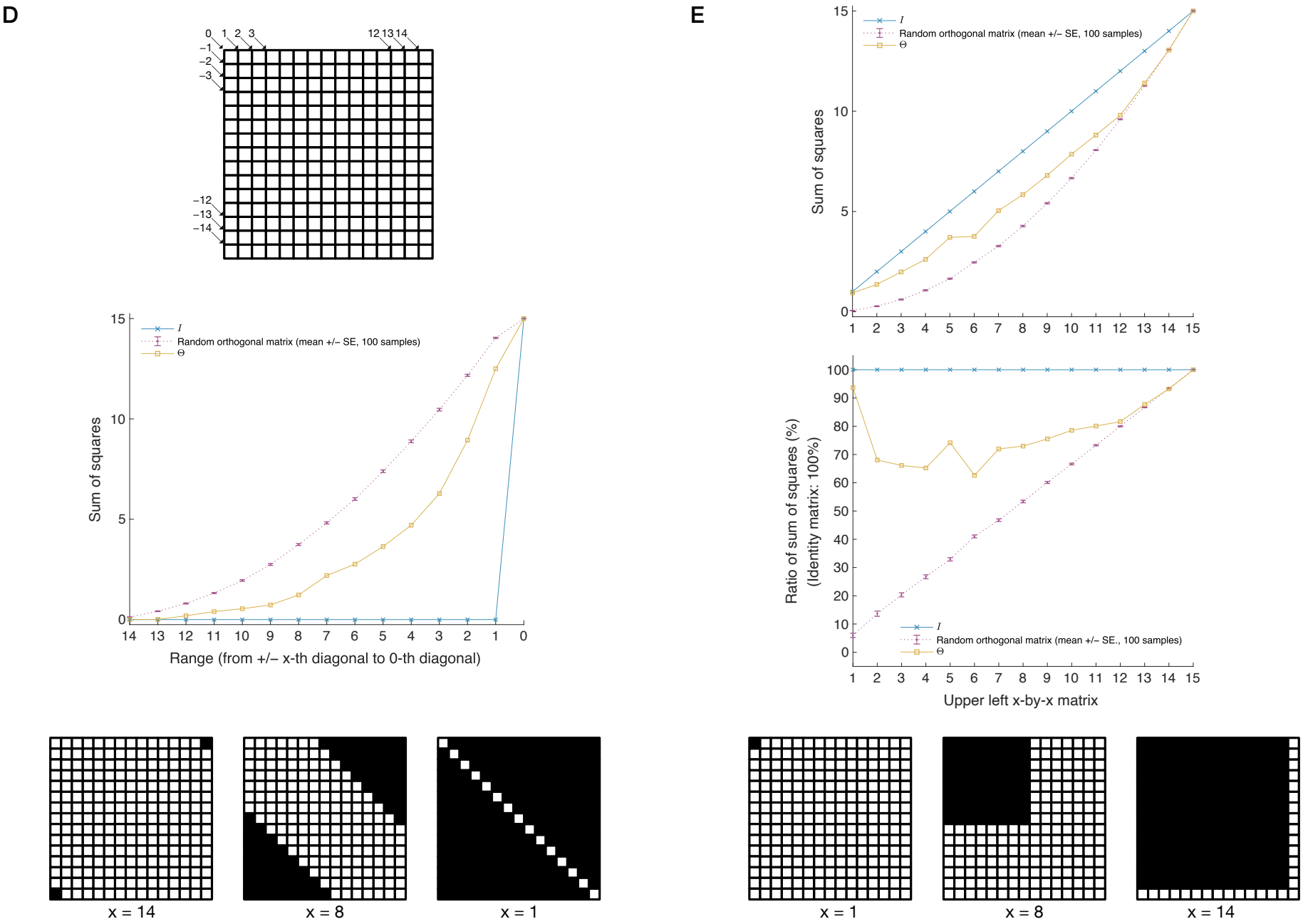
B



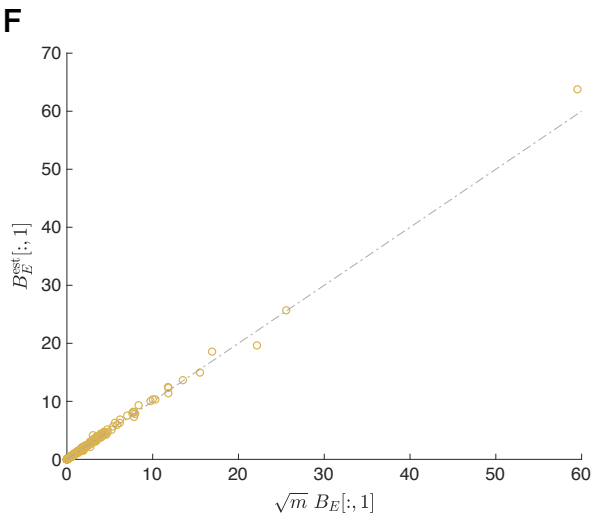
C



D



F



G

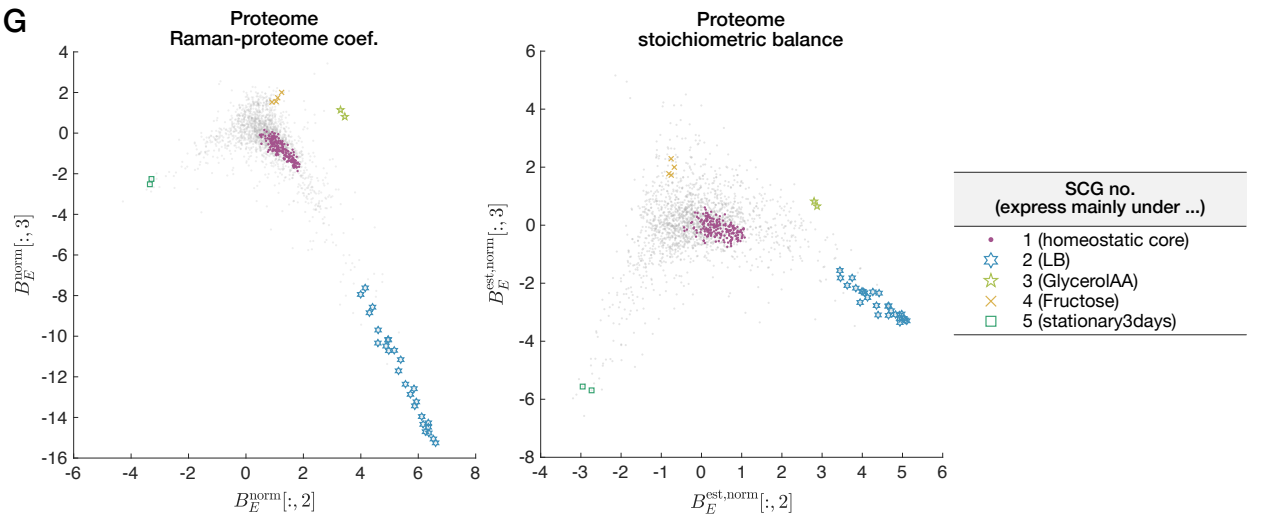


Fig. S10

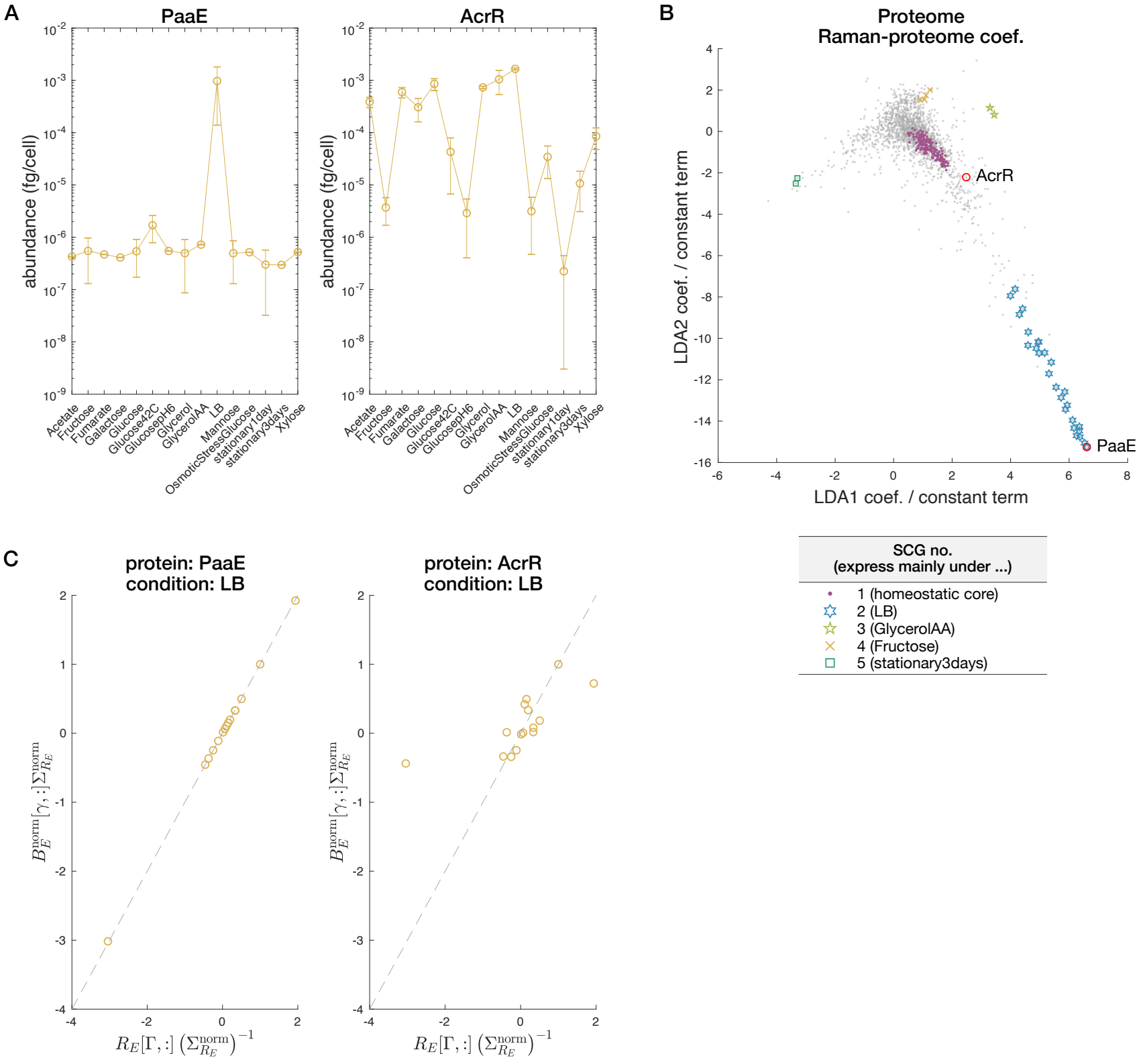
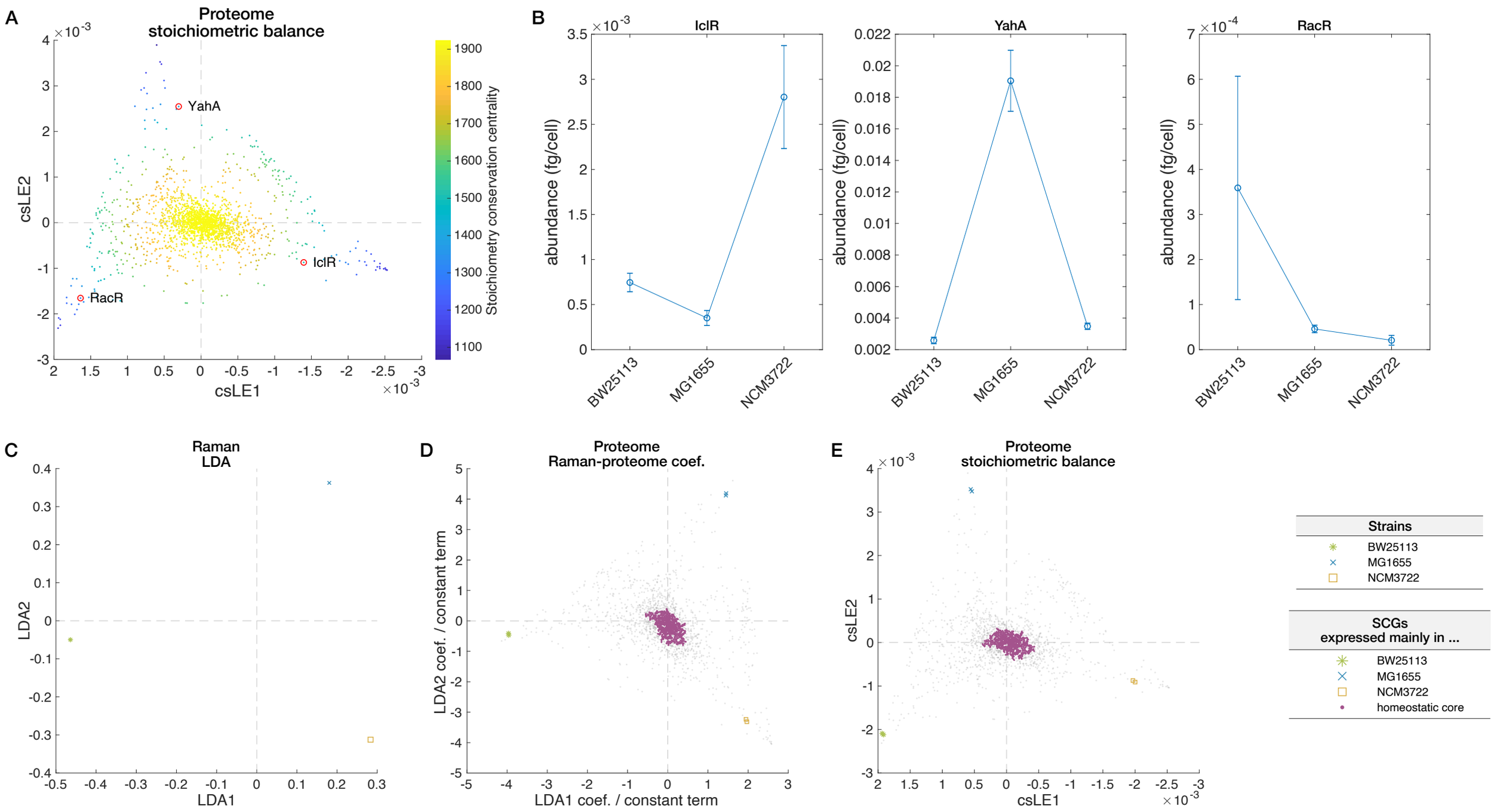
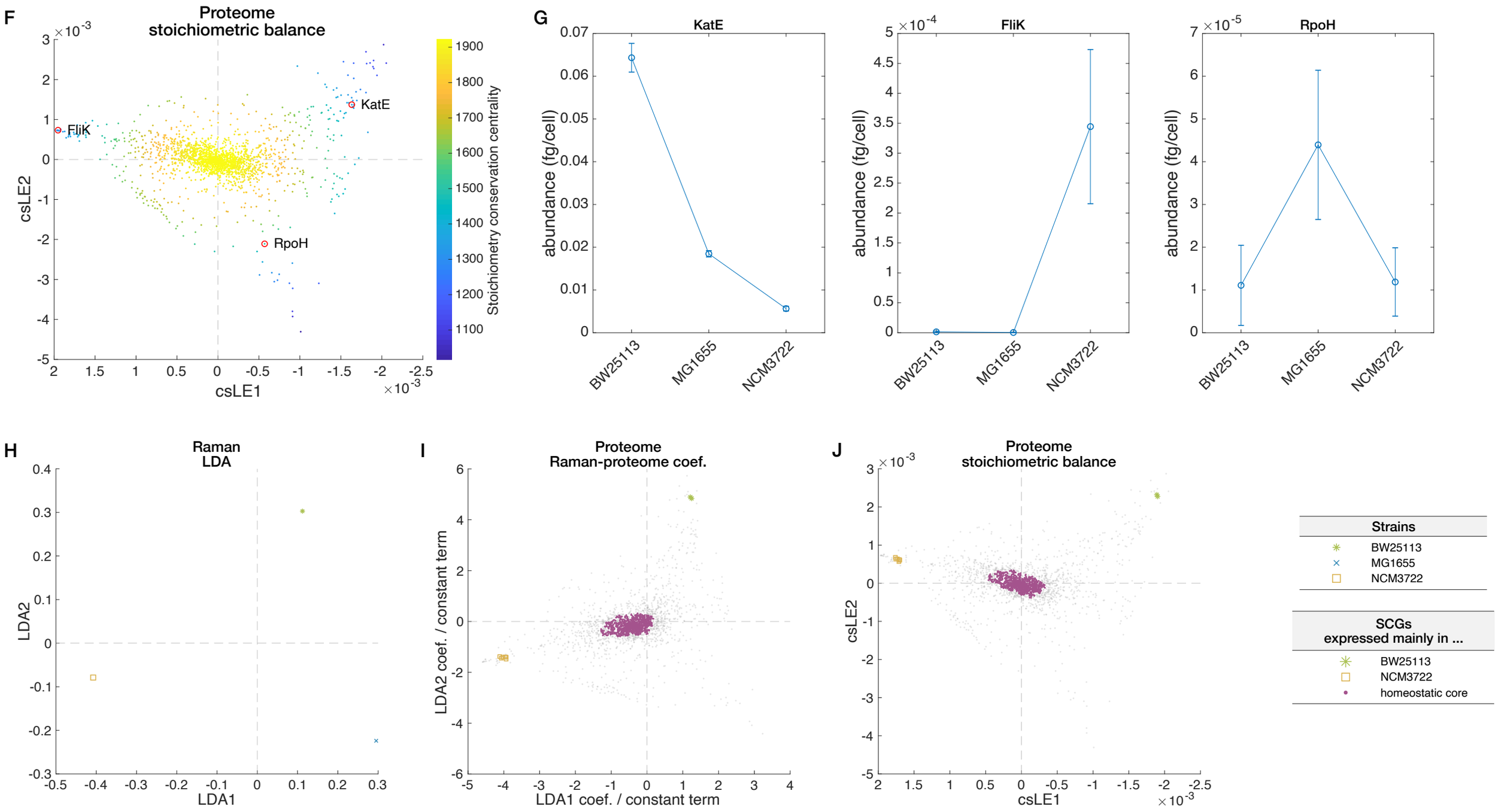

A
PaaE
AcrR
abundance (fg/cell)
Acetate
Fructose
Fumarate
Galactose
Glucose
Glucose42C
GlucosepH6
Glycerol
GlycerolAA
LB
Mannose
OsmoticStressGlucose
stationary1day
stationary3days
Xylose
B
Proteome
Raman-proteome coef.
LDA2 coef. / constant term
LDA1 coef. / constant term
AcrR
PaaE
SCG no.
(express mainly under ...)
1 (homeostatic core)
2 (LB)
3 (GlycerolAA)
4 (Fructose)
5 (stationary3days)
C
protein: PaaE
condition: LB
protein: AcrR
condition: LB
$B_E^{\mathrm{norm}}[\gamma,:]\Sigma_{R_E}^{\mathrm{norm}}$
$R_E[\Gamma,:]\left(\Sigma_{R_E}^{\mathrm{norm}}\right)^{-1}$

Fig. S11

Proteomes of *E. coli* 3 strains under LB



Proteomes of *E. coli* 3 strains under M9 Glucose



Transcriptomes of *S. pombe* under 10 conditions

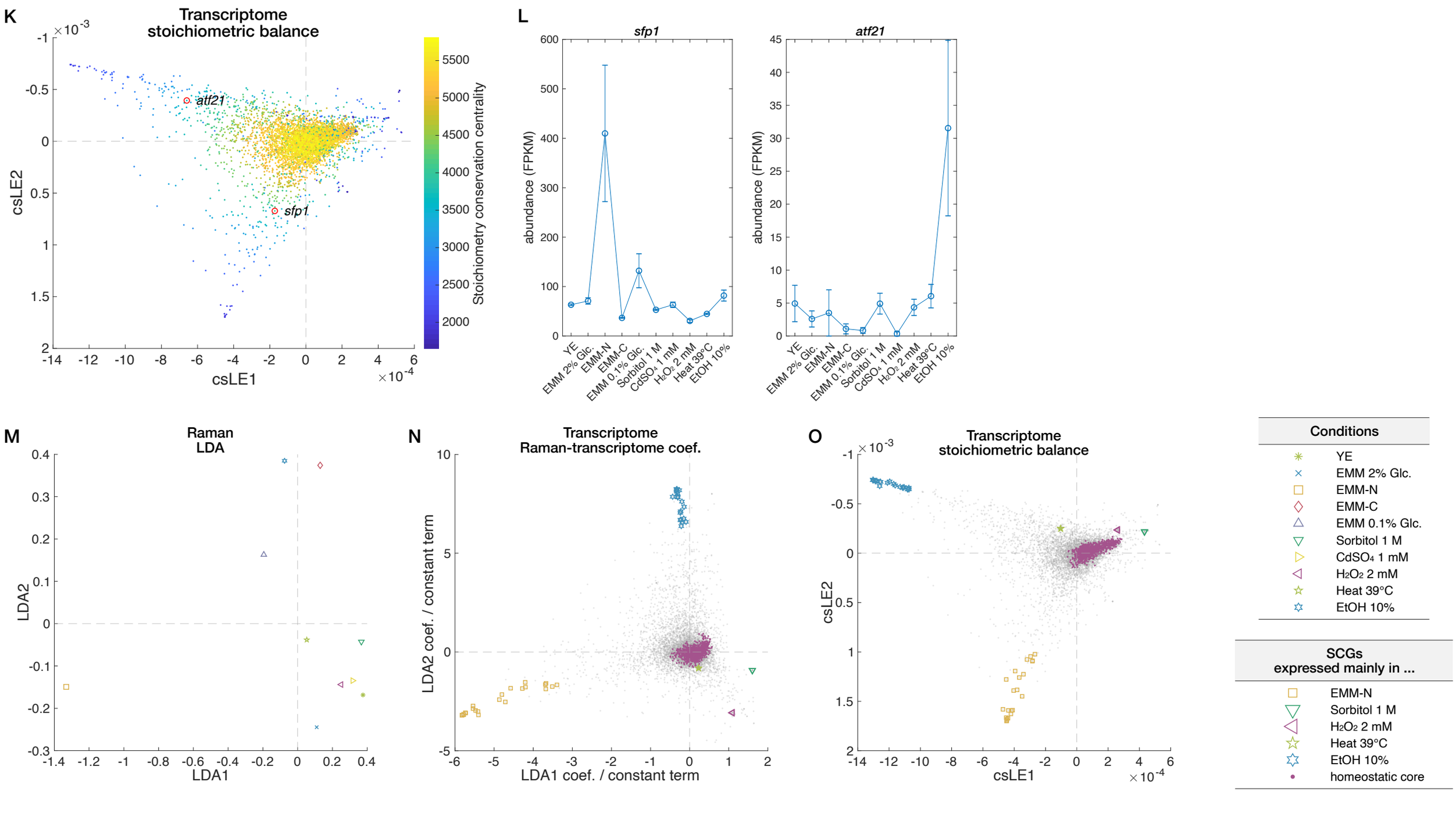
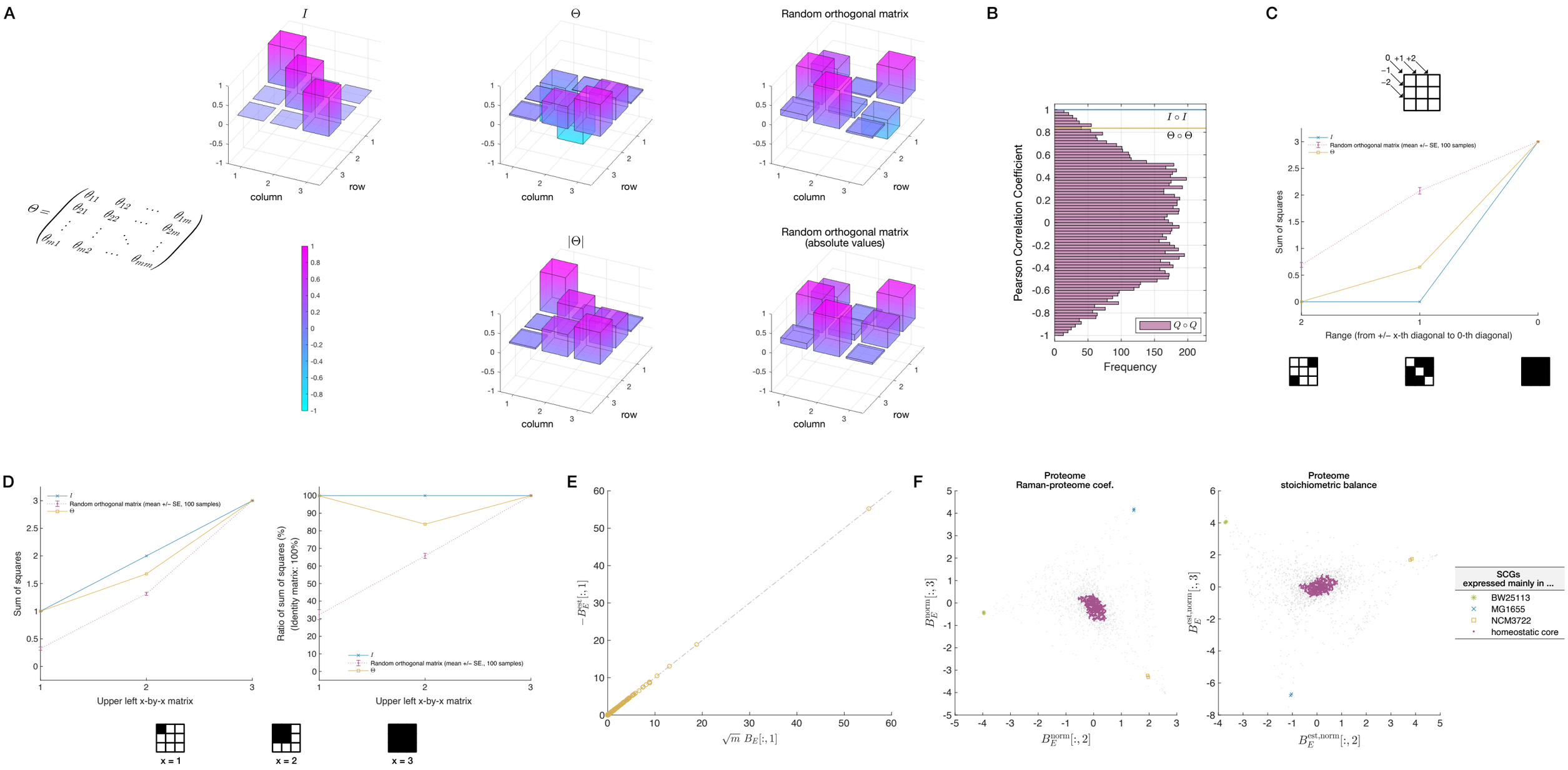
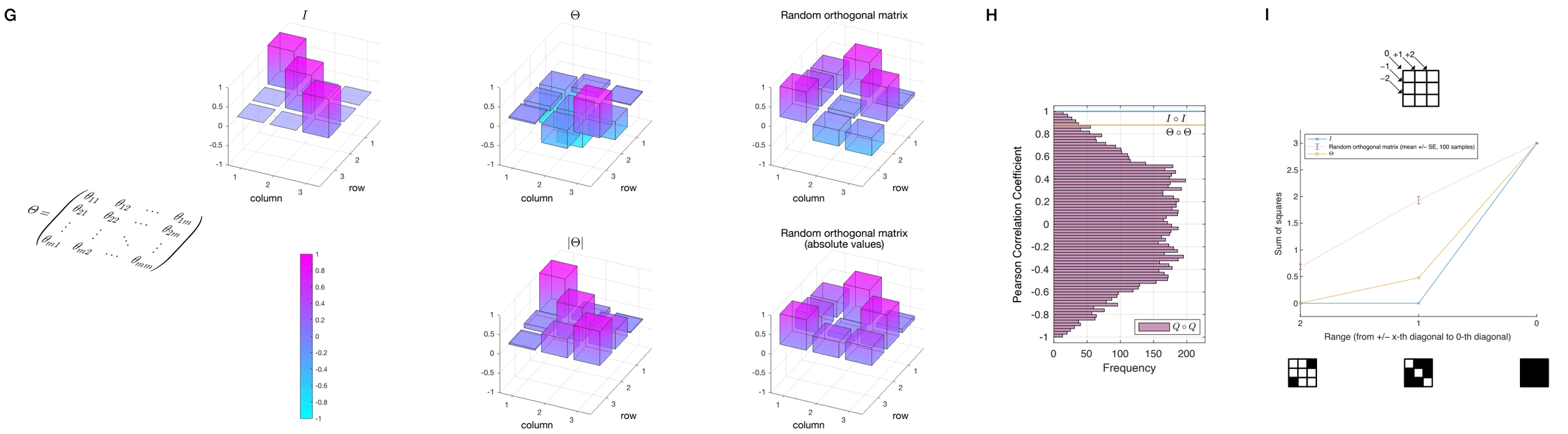


Fig. S12

Proteomes of *E. coli* 3 strains under LB



Proteomes of *E. coli* 3 strains under M9 Glucose



Transcriptomes of *S. pombe* under 10 conditions

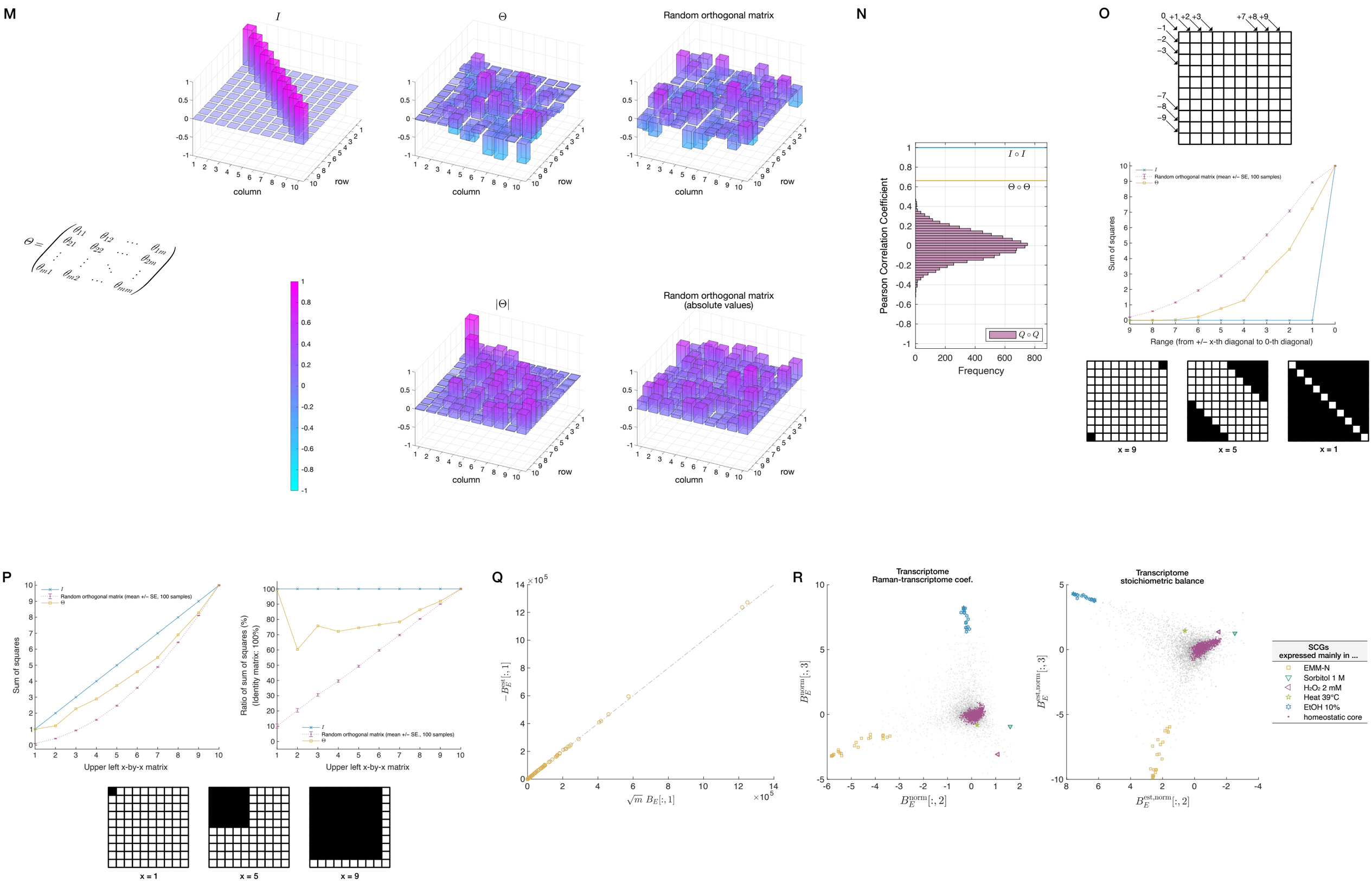
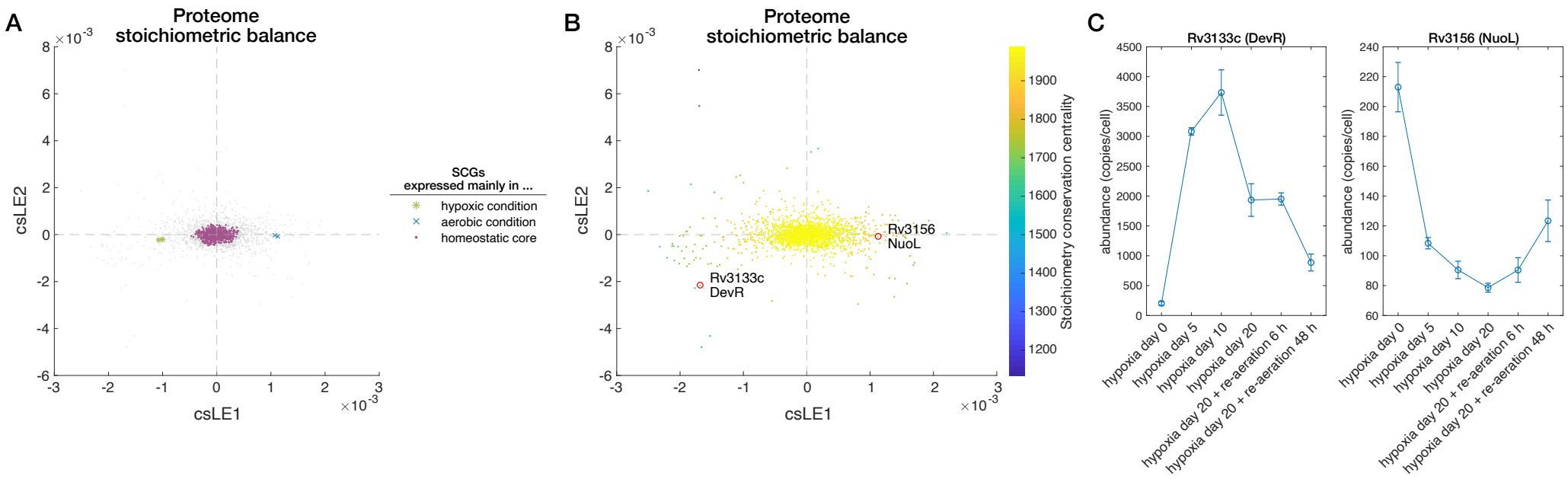
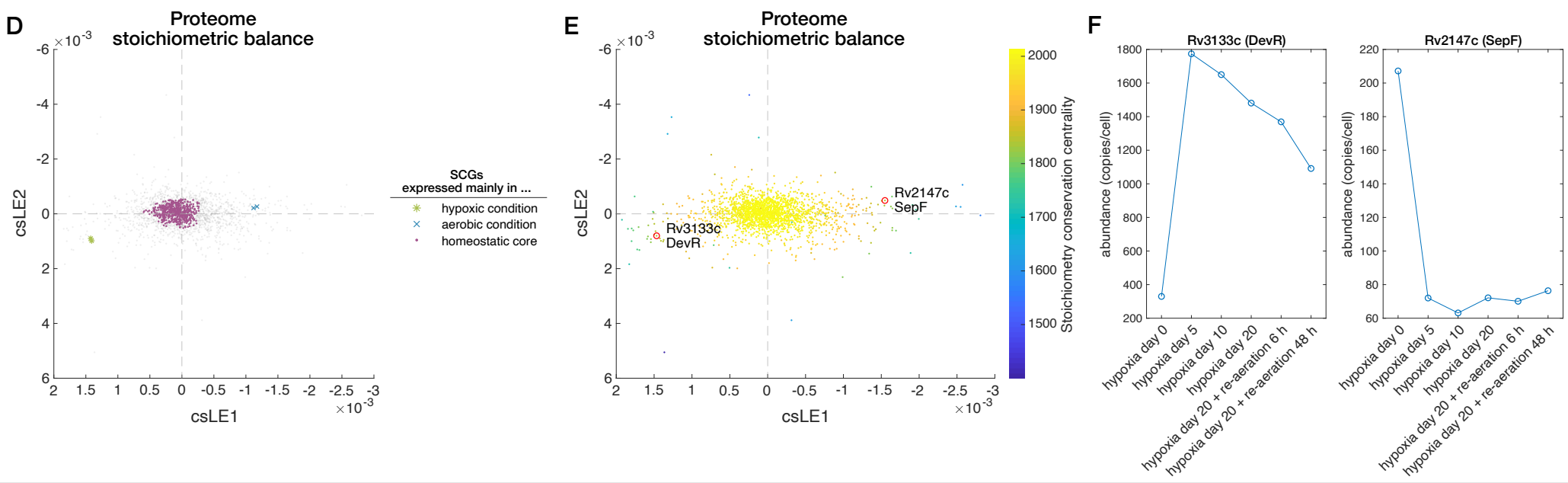


Fig. S13

Proteomes of *M. tuberculosis* H37Rv under gradual changes in oxygen levels



Proteomes of *M. bovis* BCG under gradual changes in oxygen levels



Proteomes of *S. cerevisiae* under 10 conditions in chemostat with the same dilution rate

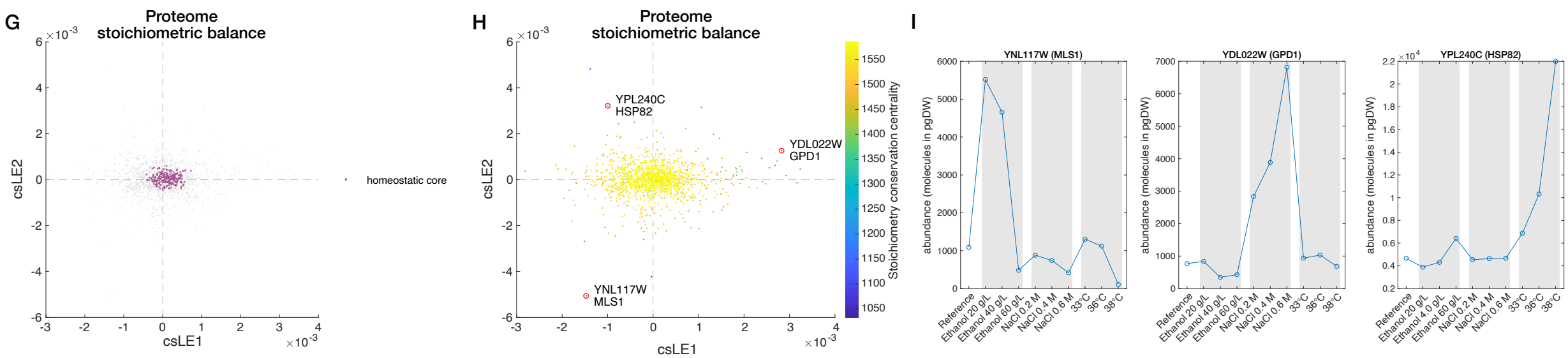


Fig. S14

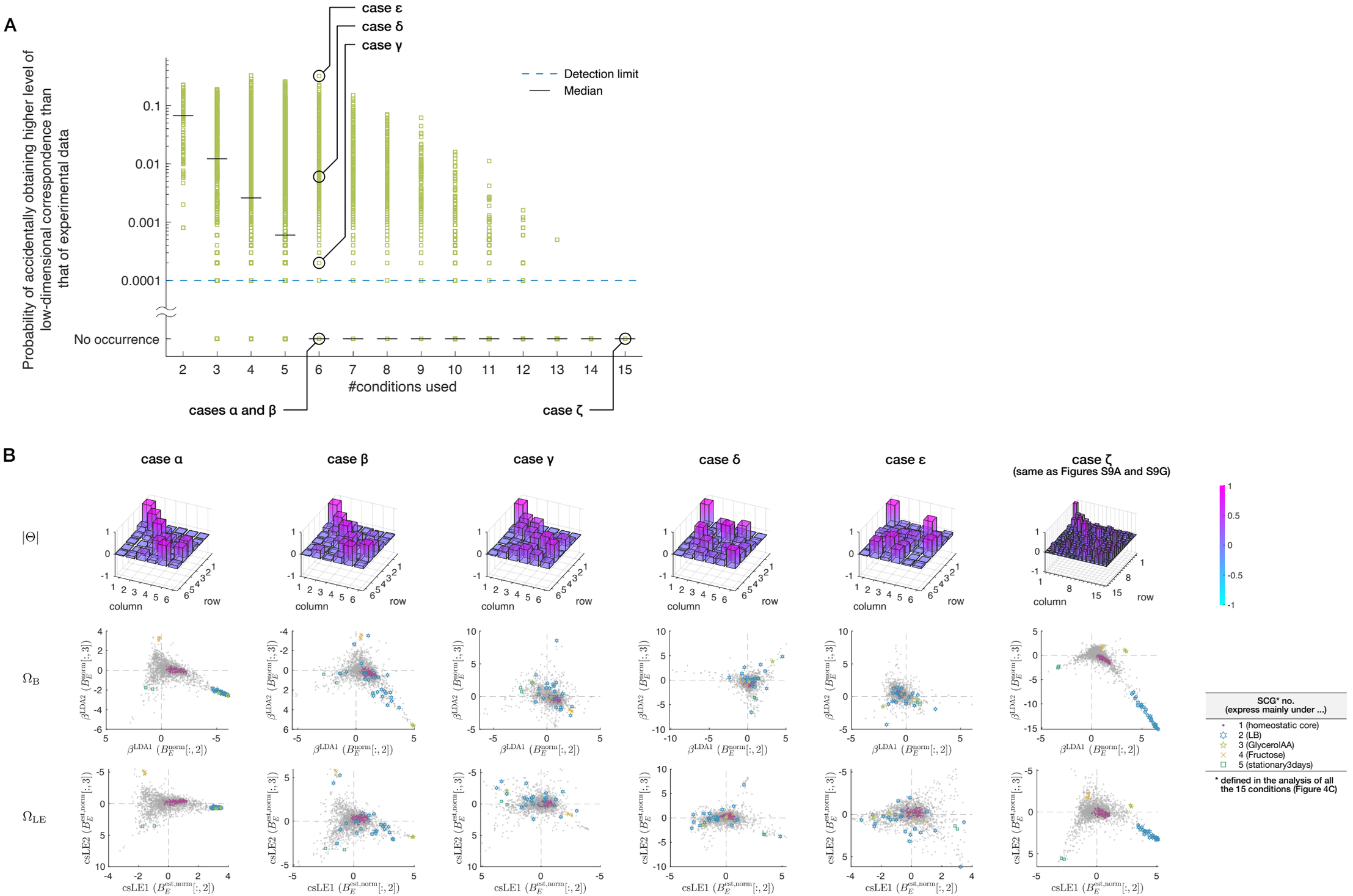

A
case ε
case δ
case γ
Detection limit
Median
Probability of accidentally obtaining higher level of low-dimensional correspondence than that of experimental data
0.1
0.01
0.001
0.0001
No occurrence
#conditions used
cases α and β
case ζ
B
case α
case β
case γ
case δ
case ε
case ζ
(same as Figures S9A and S9G)
|Θ|
column
row
Ω_B
Ω_LE
SCG* no.
(express mainly under ...)
1 (homeostatic core)
2 (LB)
3 (GlycerolAA)
4 (Fructose)
5 (stationary3days)
* defined in the analysis of all the 15 conditions (Figure 4C)

Table S1

Phase	Overview of Composition	Temperature	pH	Name in This Paper
exponential	M9 minimal media + acetate	37°C	7	Acetate
	M9 minimal media + fructose			Fructose
	M9 minimal media + fumarate			Fumarate
	M9 minimal media + galactose			Galactose
	M9 minimal media + glucose			Glucose
	M9 minimal media + glucose	42°C		Glucose42C
	M9 minimal media + glucose	37°C	6	GlucosepH6
	M9 minimal media + glycerol		7	Glycerol
	M9 minimal media + glycerol + amino acids			GlycerolAA
	M9 minimal media + glucose + NaCl			OsmoticStressGlucose
	M9 minimal media + mannose			Mannose
	M9 minimal media + xylose			Xylose
	LB			LB
stationary for 1 day	M9 minimal media + glucose			stationary1day
stationary for 3 days				stationary3days