Electronic Supplementary Material (ESI) for Molecular BioSystems.

One third of dynamic protein expression profiles can be predicted by simple rate equations

Konstantine Tchourine[1*], Christopher S. Poultney[2,3*], Li Wang[1], Gustavo M. Silva[1], Sandhya Manohar[1], Christian L. Mueller[1], Richard Bonneau[1], Christine Vogel[1]

[1]Center for Genomics and Systems Biology, New York University, New York, USA
[2]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY USA
[3]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY USA

* equally contributing authors

Supplementary material

Supplementary notes

Mathematical explanation of correlation between synthesis and degradation rates

Here, we discuss whether the correlation between synthesis (k_s) and degradation (k_d) detected in the Rapamycin data (Figure 3, main text) may be a true biological relationship or an artifact. We performed several analyses that convinced us that the observed does not reflect a biological signal. One line of evidence is that no correlation was observed between k_s and k_d rates measured in normal conditions (Table S4). Another line of evidence is that the existing correlation in the Rapamycin data is unaltered when all predicted values of k_s and k_d are taken into account, including low-confidence values (Table S5). When doing so for the other two datasets, no correlation is observed. In addition, predicted k_s and k_d values correlate in the Rapamycin data (but not the other two datasets) even if the underlying expression data was randomly shuffled (Table S5). Therefore, as the correlation is entirely specific to the Rapamycin data and also occurs in randomized versions, the effect is likely due to an artifact.

To explain this possible artifact, one has to consider the specific details of our model and of the different data normalizations performed on the three data sets. In essence, the strong correlation in the Rapamycin data is due to the fact that this data is heavily normalized (and we do not have the un-normalized data). In the case of Rapamycin data, there is a strong correlation in log space with the regression coefficient of approximately 1:

$$\log(k_s) \approx \log(k_d) + C \quad \text{(Eq S1.)}$$

for some constant C (determined to be C = -.43 using the R function lm, given log is base e; regression coeff = 1.06). This results in a different correlation in the linear space:

$$k_s \approx e^C k_d \quad \text{(Eq S2)}$$

where e^C is still a constant independent of the gene.

We may rearrange Eq. 1 (main text):

$$k_s = k_d \frac{P}{R} + \frac{\frac{dP}{dt}}{R} \quad \text{(Eq S3)}$$

This allows us to see that the linear relationship can only be achieved if two conditions hold:

(1) $\frac{dP}{dt} \ll kd * P$

(2) $\frac{P}{R} \approx const.$ w.r.t. the gene.

While (1) does not hold for all data, it holds for a portion of it, and this portion is large enough to drive the correlation for the Rapamycin data. The weak correlations found in the Diamide and Sodium chloride data sets is driven by a subset of genes that follow assumptions (1) and (2). When normalizing/transforming the Diamide and Sodium chloride data such that (2) also holds, increases the correlation between log(ks) and log(k_d) substantially (*not shown*).

Condition (2) does not necessarily hold true in general. Although protein and RNA abundances are correlated in log space (e.g. Abreu 2010), it does not imply that their ratio is constant in linear space unless their correlation coefficient is 1. This coefficient is normally not 1 because the amount of protein in a cell is normally orders of magnitude higher than the amount of corresponding RNA. But this coefficient would be 1 in case data is normalized across genes (by experiment), which is the case for the Rapamycin data. In this data set, each time point across all genes is normalized to be on average equal to 1 for both RNA and protein abundances, satisfying condition (2).

Figure S1. The two error models are correlated across all three stress data sets.

The ODE and linear error model are explained in the Methods section of the main text. Each dot represents a gene, and its two coordinates on the plot represent the Spearman correlations (Rs) between the predicted and the actual time-dependent protein concentrations using the ODE error model and the linear error model respectively. The Pearson correlations between the Rs values for diamide, sodium chloride, and rapamycin stress are 0.68, 0.78, and 0.65 respectively. More genes are predicted well using the ODE error model.

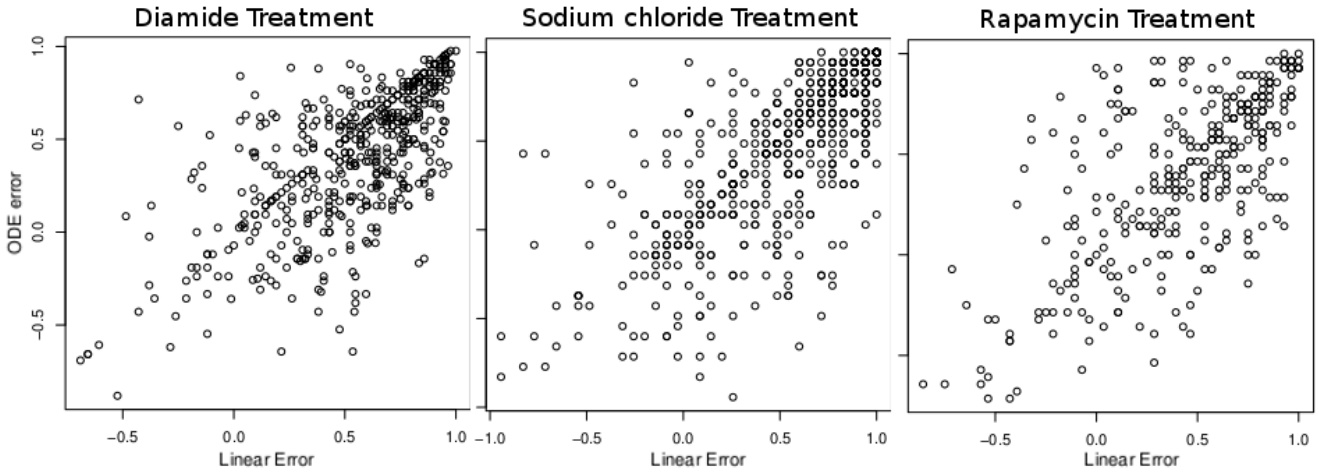


Figure S2. Error distribution

We use the ODE error model to compare the model's predictive ability for observed (blue) and randomized data (red). The vertical lines indicate the 30% FDR cutoff. Panel (A) displays the ODE error model, and (B) the linear error model (Methods). Genes are binned by the Spearman correlation between predicted and actual protein time-dependent concentration profiles. The ODE error model yields a better separation between actual data and randomly row-shuffled data than the linear error model (Table S1).

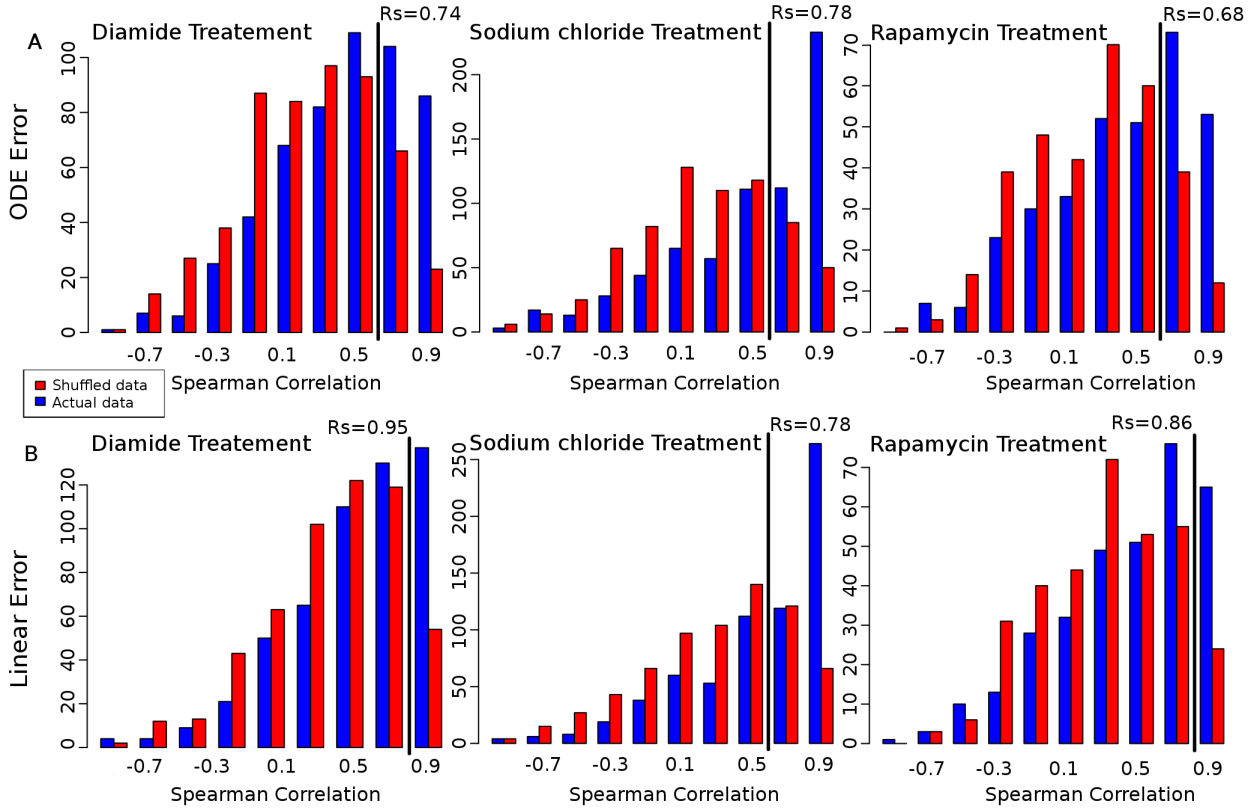


Figure S3. Average silhouette distance for profile clustering˜

Each panel shows the silhouette distance plot used to determine the optimal number of clusters for the different data sets. Although it is optimal to use 6 clusters for the sodium chloride stress, we found that removing one cluster does not change our analysis, and so 5 was chosen as the number of clusters for each data set.

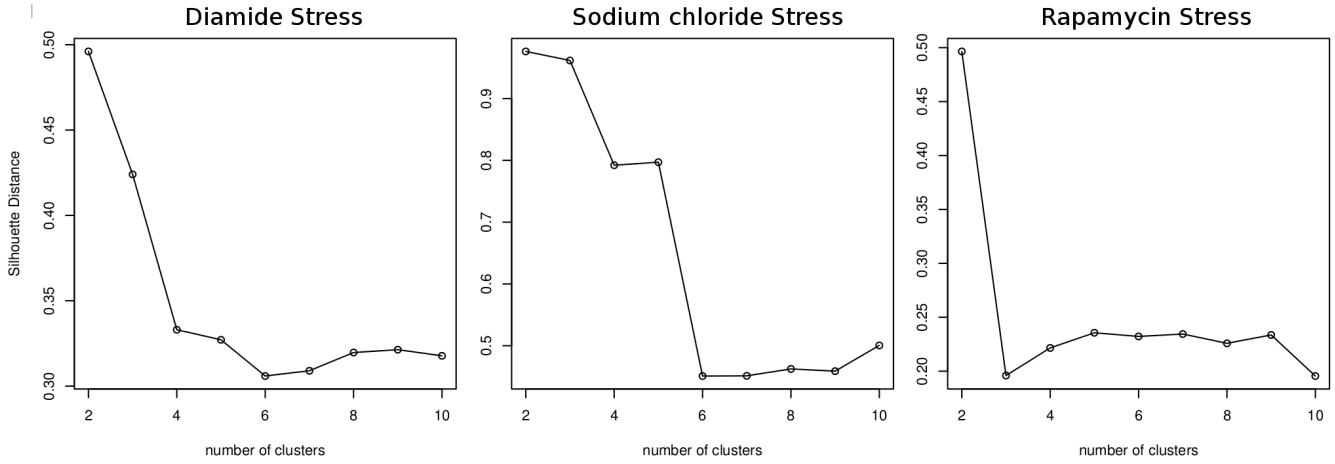


Figure S4. Differences in abundance do not account for differences in predictability

We investigated whether better prediction accuracy is achieved in highly abundant proteins whose measurements may be less noisy. No such correlation was observed. Each point represents a gene in either the Diamide (top), Sodium chloride (middle) or Rapamycin data set (bottom). Neither the median, min., nor the maximum abundance (only median abundance shown) of either the RNA (left) or the protein (right) time-series profiles is correlated with Spearman correlation between the observed and the predicted time-series protein profile. The logarithms presented are natural logs.

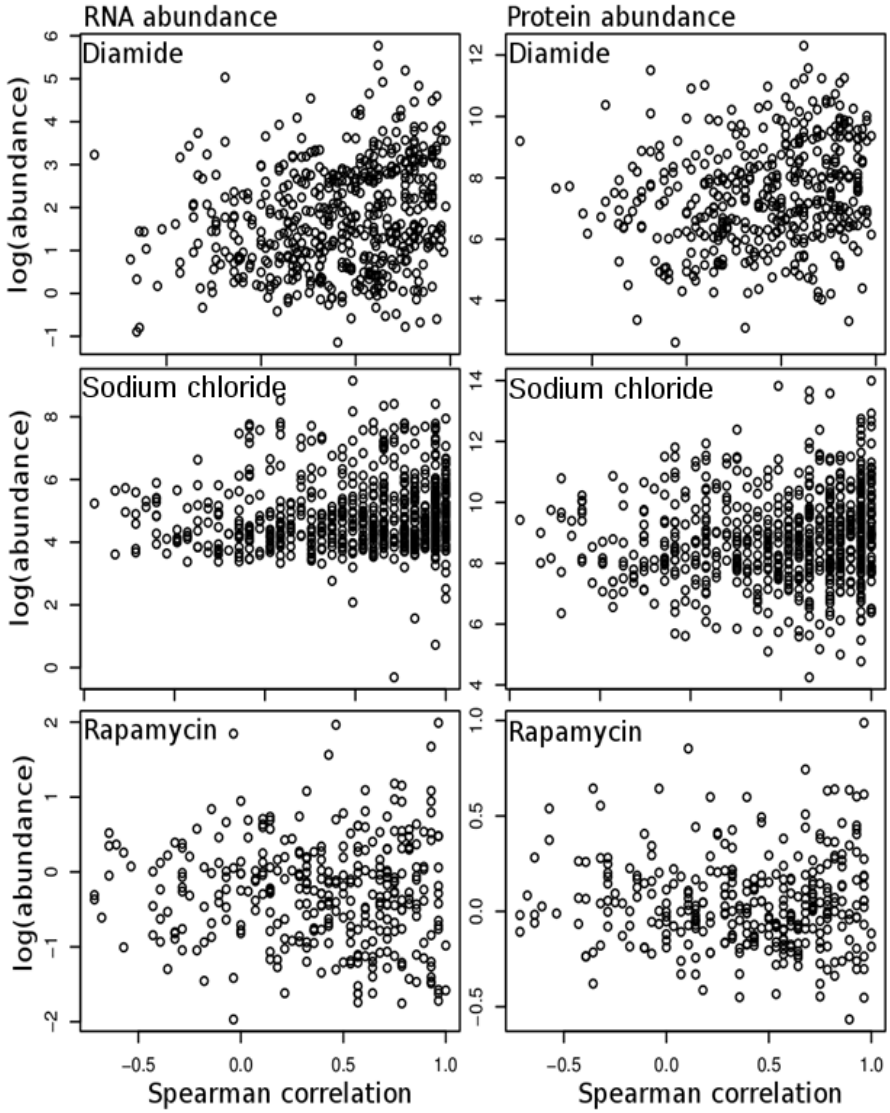


Figure S5. All relative expression profiles clustered by profile and by parameter landscape.

For each stress, the expression profiles are represented by a heat map. Each row corresponds to a gene, and the genes are ordered by the hierarchical clustering of their expression profiles. Black horizontal lines denote the boundaries between the 5 clusters. For mRNA and protein expression columns, the color at each time point represents the expression level of the gene's RNA and protein at the specified time point relative to the 0 time point (natural log-ratio, see color bar on the right; 0 time points are omitted). The rightmost column for each stress denotes the Parameter landscape (PL) cluster of each well-predicted gene. The color-cluster correspondence is outlined on the color bar on the right, with the numbering consistent with Figure S8.

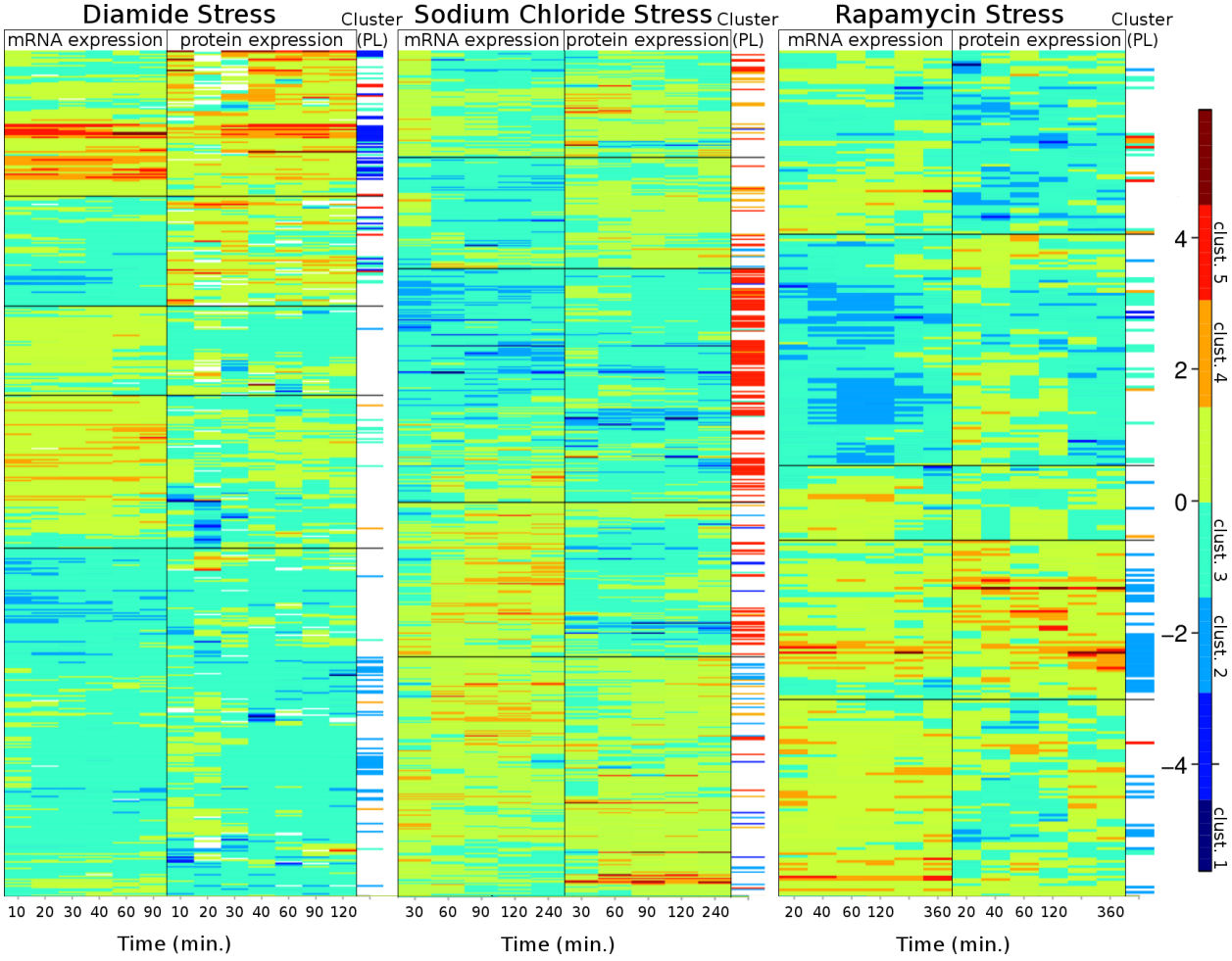


Figure S6. Median profiles of each profile cluster

The median profile for each profile cluster for each stress data set is plotted below. The black line on each panel corresponds to the protein abundance at every time point (relative to the first time point, in terms of natural log-ratio; P(t)), with the dots representing the median across the cluster and with the error bars spanning 68% of the values in each cluster, and the red line, dots and error bars represent the same statistics for the relative RNA abundances (R(t)). Number of genes in each cluster is shown on the top right corner of every panel.

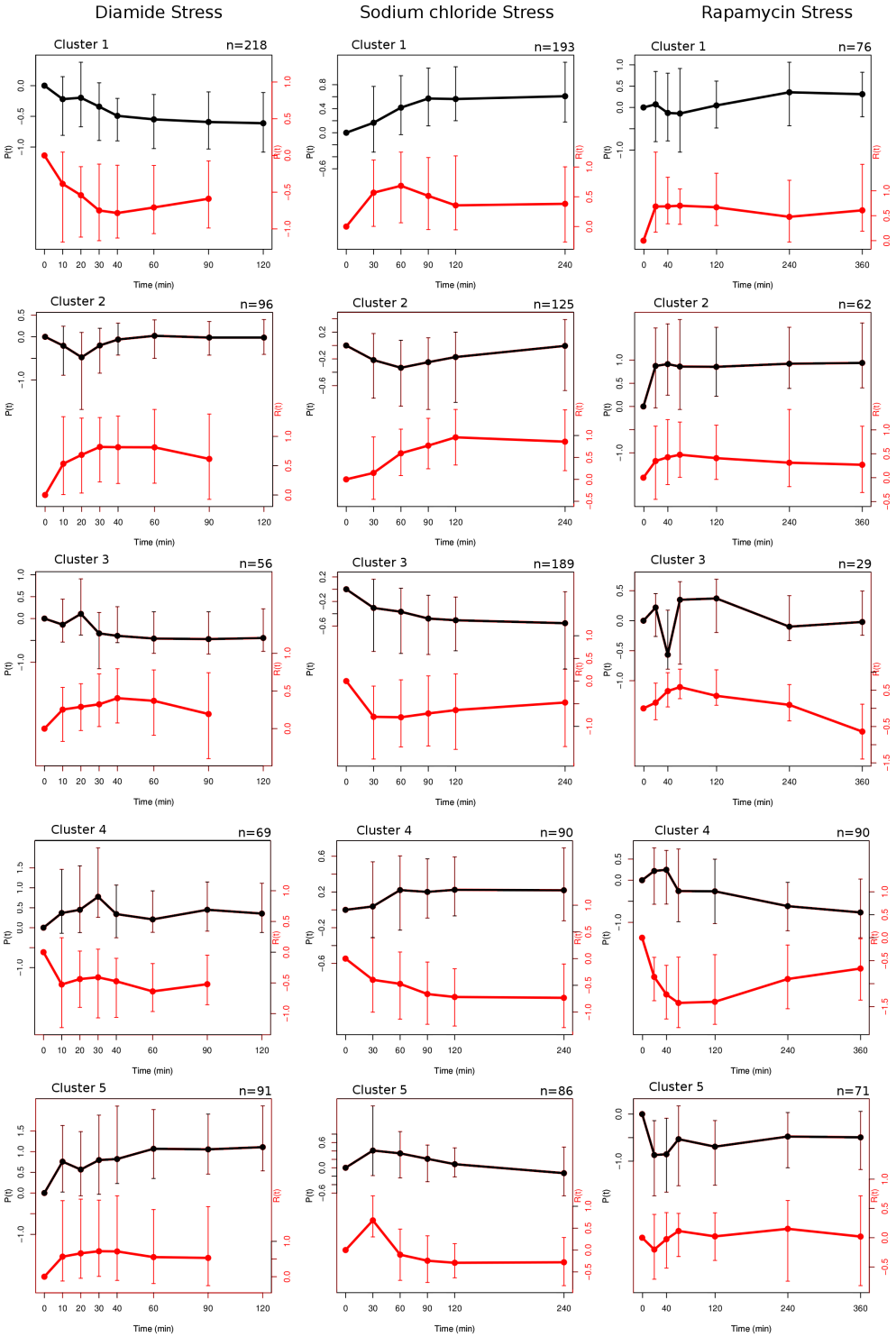


Figure S7. Comparison of measures of prediction accuracy

Each point represents a gene, with its coordinates representing the correlation between the gene's predicted protein levels and observed protein abundances. Each x-coordinate represents the Spearman correlation between the two concentrations, while each y-coordinate represents the fraction of variance unexplained between observed and predicted protein abundances. The two measures correlate. Spearman correlation (Rs) is used as a measure of prediction accuracy (main text). Rs values are displayed in the top right corner of each panel.

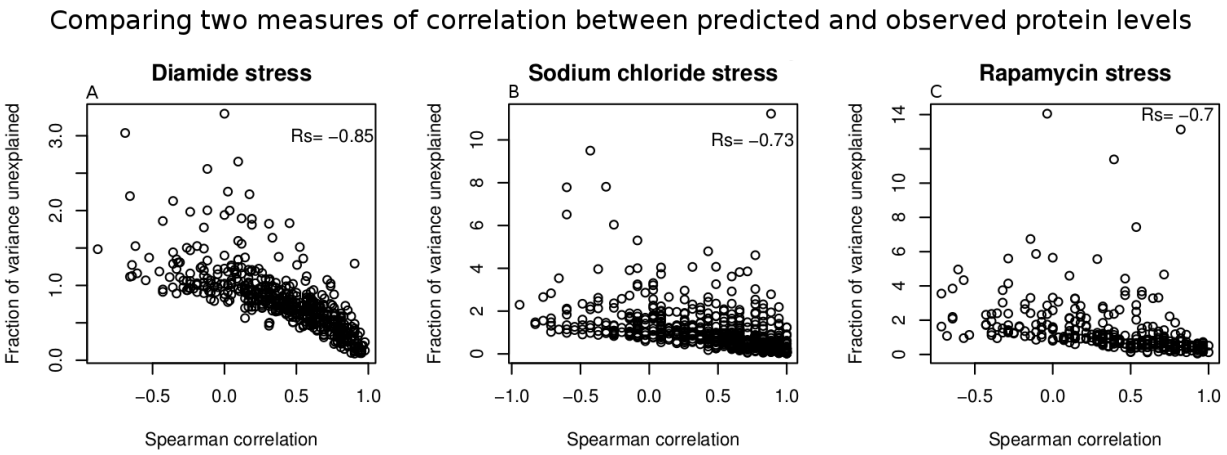


Figure S8. Representative profiles for each cluster of parameter landscapes

Each panel shows a representative parameter landscape (PL) for each of the 5 PL clusters in the three data sets. The representative PLs were picked manually. The black circle on each heatmap represents the parameters that optimize the prediction (reported k_s and k_d values), and the white circle marks the center of the heat map.

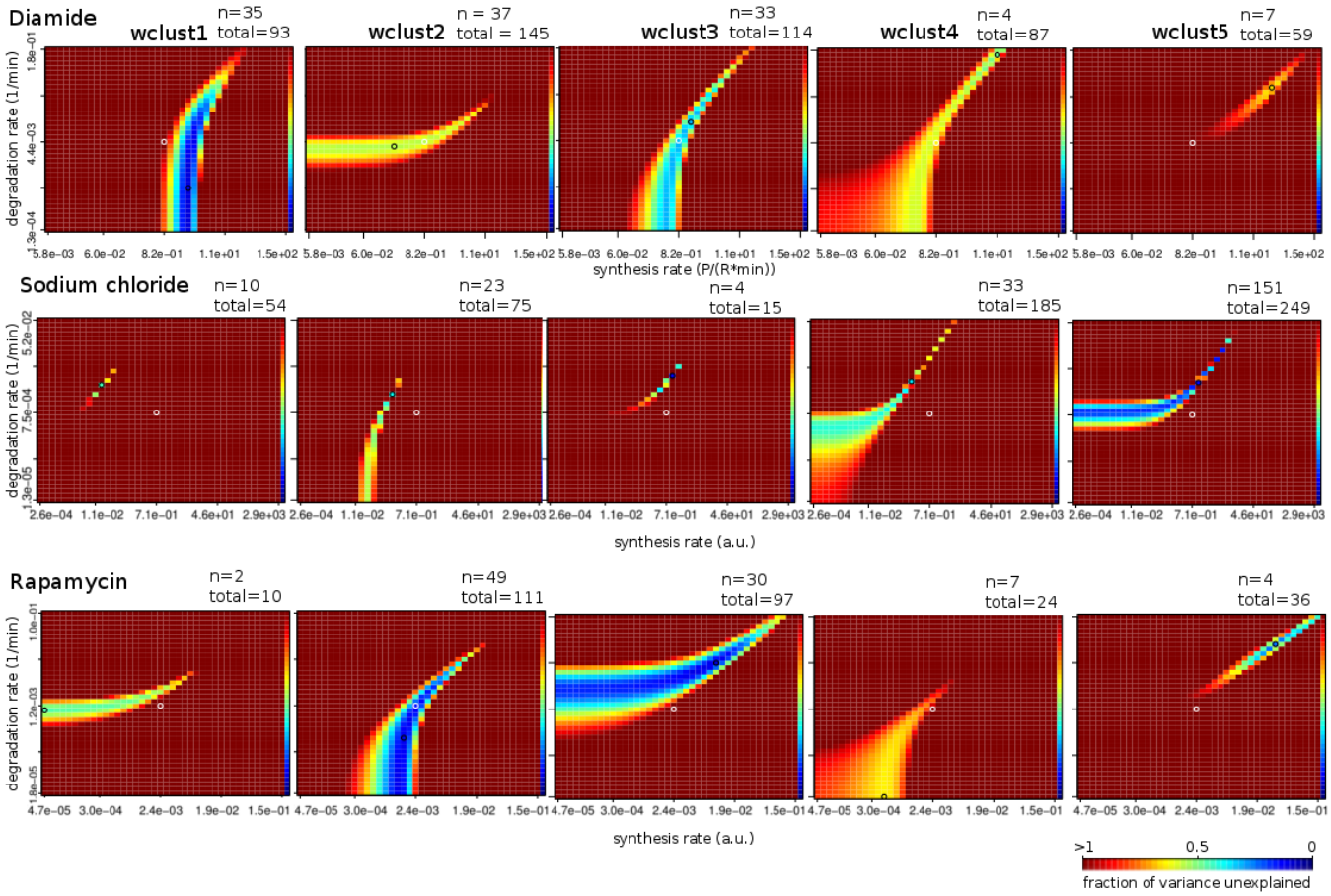


Figure S9. Distribution of predicted synthesis and degradation rates by cluster.

The coordinates of each point are the k_s and k_d values estimated per protein (<30% FDR). Different profile clusters (as in Figure S5, S6) are denoted by different colors/shapes. No separation by cluster is apparent. Rates in the Sodium chloride and Rapamycin data set are presented in arbitrary units (a. u.), while Diamide Stress data set rates are expressed in absolute protein concentration (P) and absolute mRNA concentration (R) in number of molecules per cell.

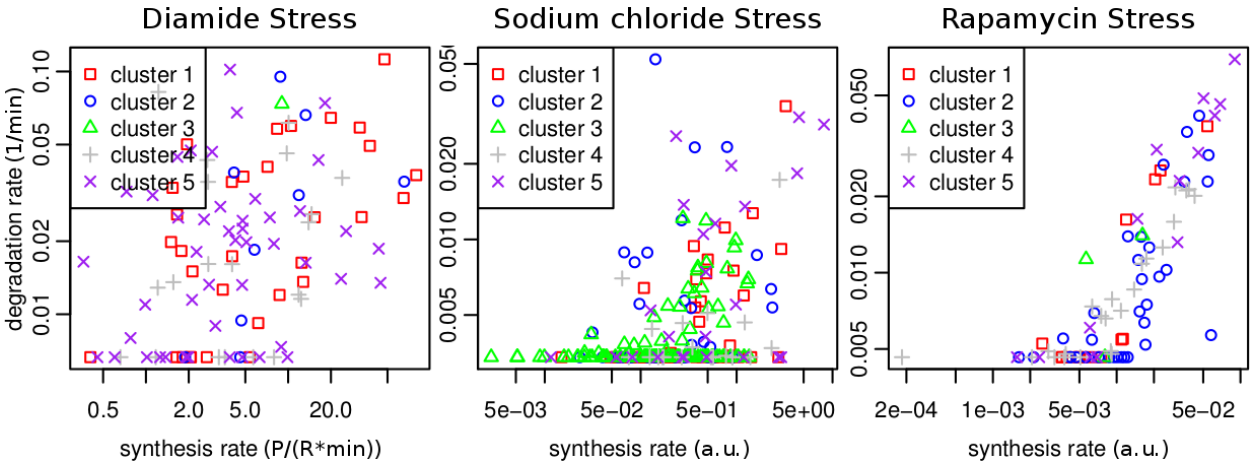
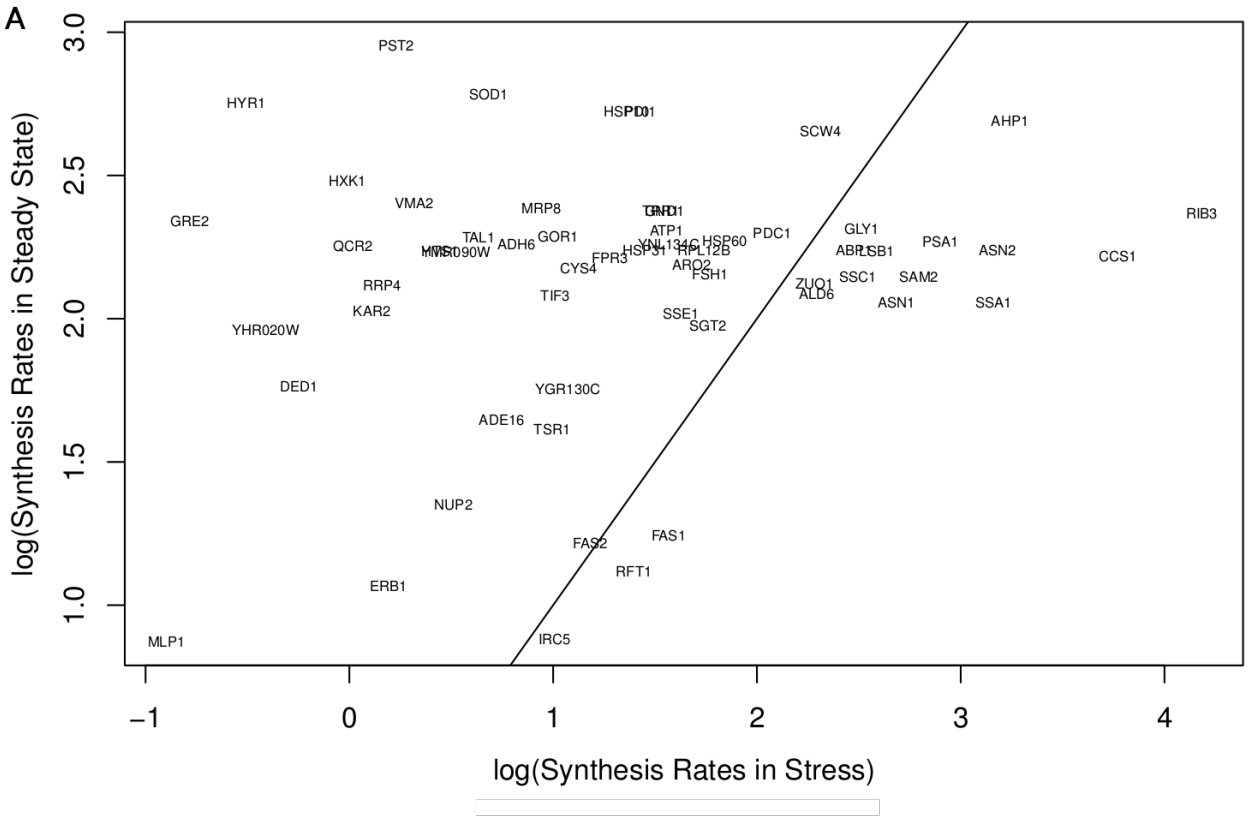


Figure S10. Gene-by-gene comparison between measured and fitted rates

Highest-confidence predictions for synthesis and degradation rates in diamide stress conditions were compared on a gene-by-gene basis to the corresponding experimentally measured rates obtained in steady state conditions. Panel (A) plots the synthesis rates of the genes that are predicted well (i.e. 30% FDR) according to our model and whose responses to stress are degradation-independent according to parameter landscape analysis. Panel (B) plots the degradation rates of the well-predicted genes with synthesis-independent parameter landscape profiles. In both cases, the parameter landscape restriction selects the genes with the highest confidences of predictions of either synthesis or degradation rates. The diagonal line shows where the predicted and observed rates are exactly equal. Logarithms are taken in base *e*.


A
log(Synthesis Rates in Steady State)
log(Synthesis Rates in Stress)
PST2
SOD1
HYR1
AHP1
SCW4
HXK1
VMA2
MRP8
GRE2
RIB3
QCR2
TAL1
ADH6
GOR1
PDC1
GLY1
PSA1
ASN2
CCS1
RRP4
CYS4
FSH1
SSC1
SAM2
ZUO1
ALD6
ASN1
SSA1
KAR2
TIF3
SSE1
SGT2
YHR020W
DED1
YGR130C
ADE16
TSR1
NUP2
FAS2
FAS1
RFT1
ERB1
MLP1
IRC5

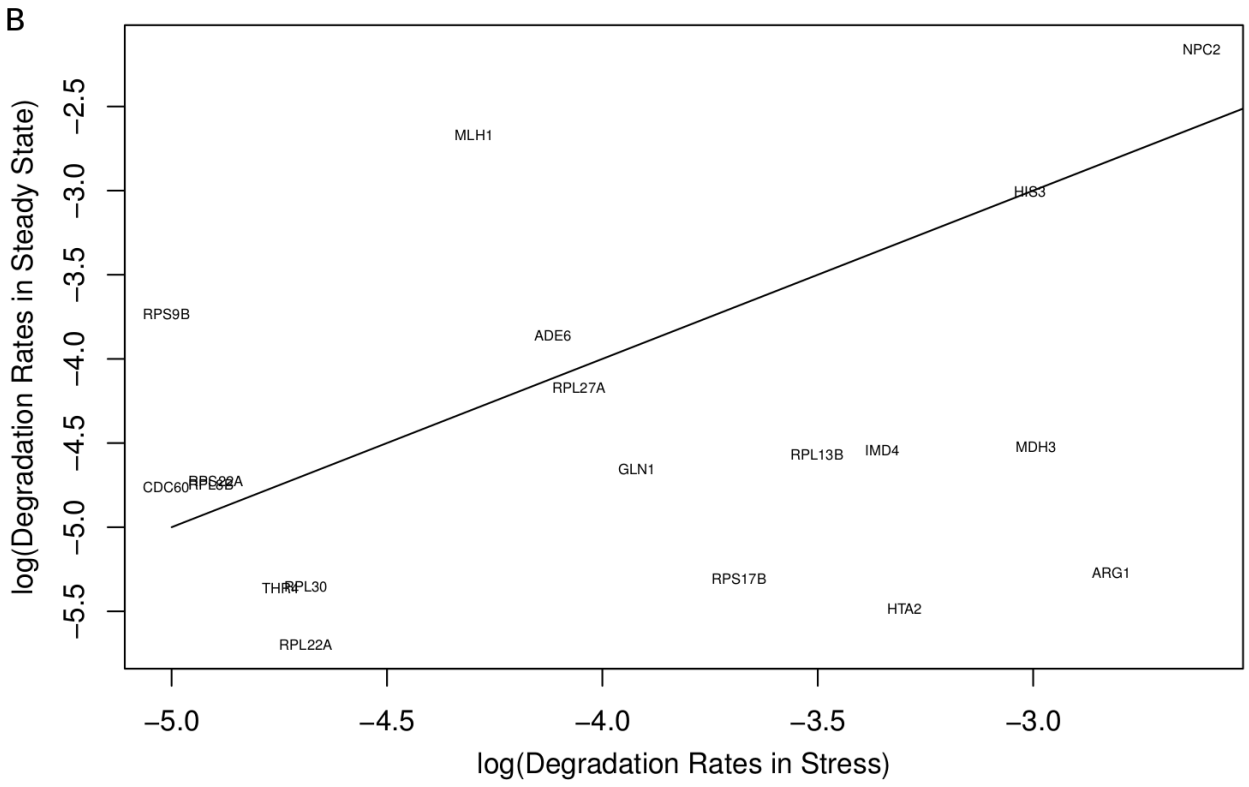

B
log(Degradation Rates in Steady State)
log(Degradation Rates in Stress)
NPC2
MLH1
HIS3
RPS9B
ADE6
RPL27A
RPL13B
IMD4
MDH3
GLN1
CDC60
RPS17B
ARG1
THR4
RPL30
HTA2
RPL22A

Table S1.

Shown are: Minimum Spearman correlation (Rs) values necessary to satisfy the 30% FDR cutoff as obtained from Figure S2, number of genes satisfying the FDR/Rs cutoff, for either the ODE model or the linear model (see Methods). For the S. pombe dataset, no Rs cutoff was found to fulfill the 30% FDR requirement in the linear error model.

	Diamide treatment	**Sodium chloride treatment**	**Rapamycin treatment**	**S. pombe oxidative stress**
Rs (ODE model)	0.74	0.78	0.68	0.83
# of genes (ODE model)	116	233	94	248
Rs (linear model)	0.95	0.78	0.86	N/A
# of genes (linear model)	24	264	33	N/A

Table S2.

Shown are: Proportion of well-predicted genes (FDR<30%) by cluster in each data set. Last column shows the proportion of well-predicted genes within the entire data set, for each data set. In parenthesis, n shows the total number of genes within each category. Highest enrichments are highlighted in **bold**, while the exceptionally low proportions are in *italic*.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Diamide	0.18 (n = 218)	*0.094* (n = 96)	*0.018* (n = 56)	0.29 (n = 69)	**0.51** (n = 91)	0.22 (n = 530)
Sodium chloride	0.22 (n = 193)	0.27 (n = 125)	**0.59** (n = 189)	0.23 (n = 90)	0.28 (n = 86)	0.34 (n = 683)
Rapamycin)	0.18 (n = 76)	**0.56** (n = 62)	*0.10* (n = 29)	0.28 (n = 90)	0.24 (n = 71)	0.29 (n = 328)

Table S3.

Shown are: Number of genes in the intersections. Clust denotes profile clusters, wclust denotes parameter landscape clusters. Numbers in the parentheses indicate the number of genes in the specified intersection that are predicted with FDR<30%. Pearson's Chi-square test detects a contingency between the two variables (clust and wclust) using the FDR<30%, with p<2e-16 for diamide and sodium chloride stresses, and p=8e-13 for rapamycin stress.

Diamide	Clust1	Clust2	Clust3	Clust4	Clust5	Total
Total	218(40)	96(9)	56(1)	69(20)	91(46)	530(116)
wclust1	5(0)	13(0)	5(0)	20(6)	50(29)	93(35)
wclust2	119(36)	9(0)	17(1)	0(0)	0(0)	145(37)
wclust3	14(2)	32(7)	6(0)	30(11)	32(13)	114(33)
wclust4	36(2)	31(2)	17(0)	3(0)	0(0)	87(4)
wclust5	23(0)	9(0)	6(0)	14(3)	7(4)	59(7)
colSum	197(40)	94(9)	51(1)	67(20)	89(46)	498(116)

Sodium chloride	Clust1	Clust2	Clust3	Clust4	Clust5	Total
total	193	125	189	90	86	683(233)
wclust1	14(4)	26(4)	11(1)	1(0)	2(1)	54(10)
wclust2	59(18)	4(1)	1(0)	9(4)	2(0)	75(23)
wclust3	4(0)	5(1)	5(3)	1(0)	0(0)	15(4)
wclust4	66(8)	10(2)	13(3)	44(9)	52(11)	185(33)
wclust5	17(6)	54(25)	134(104)	23(7)	21(9)	249(151)
colSum	160(36)	99(33)	164(111)	78(20)	77(21)	578(221)

Rapamycin	Clust1	Clust2	Clust3	Clust4	Clust5	Total
total	76	62	29	90	71	328
wclust1	0(0)	0(0)	1(0)	5(2)	4(0)	10(2)
wclust2	45(11)	43(34)	11(2)	6(1)	6(1)	111(49)
wclust3	10(2)	0(0)	3(0)	54(19)	30(9)	97(30)
wclust4	6(0)	2(0)	7(1)	5(2)	4(4)	24(7)
wclust5	3(1)	8(0)	6(0)	8(0)	11(3)	36(4)
colSum	64(14)	53(34)	28(3)	78(24)	55(17)	278(92)

Table S4.

Steady-state rates data does not exhibit a correlation between rates of synthesis and degradation. Shown are the correlations between the synthesis and degradation rates measured in steady state (Ref. Belle 2006, Fraser 2004) for genes also found in each of the three stress data sets. For each data set, neither the set of all genes nor the set of well-predicted genes (FDR 30%) exhibits a correlation between the rates of synthesis and degradation when measured in the steady state.

Intersections with data sets:	**Diamide**	**Sodium chloride**	**Rapamycin**	**All steady-state**
All data	-0.29	-0.23	-0.18	-0.25
FDR 30%	-0.25	-0.21	-0.27	

Table S5.

Pearson correlations between rates of synthesis and degradation. Rates derived from randomly shuffled time-dependent RNA and protein data yield approximately the same correlations as actual time-series data from all genes and from the best-predicted genes (FDR 30%).

	Diamide treatment	**Sodium chloride treatment**	**Rapamycin treatment**
All data	0.40	0.43	0.74
FDR 30%	0.45	0.38	0.80
Randomly shuffled data	0.41	0.49	0.76

Table S6.

List of sources of the data for the 35 static sequence features that were tested for associations with predictability and trends in synthesis and degradation rates. The results of each association test can be found in the Supplementary Data.

Data type	Source
Menadione, ribosome association (log2 stress/control)	Halbeisen, 2009
Molecular weight, PI, CAI, protein length, codon bias (CBI), relative amino acid frequencies, FOP score, GRAVY score, and AROMATICITY score	*Saccharomyces* Genome Database (Cherry, 1998)
PEST protein degradation signal	Max. score in ePESTfind (Rice, 2000)
DISEMBLE length, DISEMBL coils and hot loops (absolute and relative)	Measure of disorder of protein, which is associated with stability (Linding 2003)

Table S7.

This table includes all significant associations detected between the 35 sequence features listed in Table S6 and either prediction accuracy or synthesis or degradation rates. The analysis for rates was only done using the Diamide stress data set, where the predicted rates were realistic. The p-values shown below are not corrected for multiple hypothesis testing. For the prediction accuracy association analysis, the Wilcoxon test was used to compare the distribution of codon usage bias scores in the set of well-predicted genes to the corresponding distribution among the same number of genes that had the lowest prediction quality scores (i.e. Spearman correlation between observed and predicted profiles). For the association with rates, the Wilcoxon test was performed by comparing the distribution of the codon bias scores of the top quarter of well-predicted genes (with highest predicted synthesis or degradation rates) to the corresponding distribution among the bottom quarter of well-predicted genes.

	Association with prediction quality						Assoc. with rates (Diamide stress)			
	Pearson Correlation			Wilcoxon test (p-value)			Pearson correlation		Wilcoxon (p-val.)	
	Diamide	NaCl	Rap.	Diamide	NaCl	Rap.	Synth.	Degr.	Synth	Degr.
CAI	0.16	0.17	0.04	8e-04	2e-04	0.5	-0.10	-0.31	0.28	2.5e-03
CBI	0.16	0.13	0.04	5e-04	5e-04	0.5	-0.07	-0.31	0.42	2.8e-03
FOP	0.16	0.14	0.04	5e-04	2e-04	0.5	-0.08	-0.32	0.36	2.3e-03

References

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26, 73–79

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009 Dec;5(12):1512-26.

Halbeisen, Regula E., and André P. Gerber. "Stress-dependent coordination of transcriptome and translatome in yeast." PLoS biology 7.5 (2009): e1000105.

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: Implications for structural proteomics. Structure 11, 1453–1459

Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276 –277