Stochastic protein multimerization, activity, and fitness

Kyle Hagner, Sima Setayeshgar, and Michael Lynch

Phys. Rev. E **98**, 062401 — Published 7 December 2018

DOI: 10.1103/PhysRevE.98.062401

Stochastic Protein Multimerization, Activity and Fitness

Kyle Hagner,[1] Sima Setayeshgar,[1] and Michael Lynch[2]

[1]*Department of Physics, Indiana University, Bloomington, Indiana 47405*
[2]*Biodesign Center for Mechanisms of Evolution,*
Arizona State University, Tempe, Arizona 85287[*]

Many proteins assemble into homo-multimeric structures, with a number of subunits that can vary substantially among phylogenetic lineages. As protein-protein interactions require productive encounters among subunits, such variation might partially be explained by variation in cellular protein abundance. Protein abundance in turn depends on the intrinsic rates of production and decay of mRNA and protein molecules, as well as rates of cell growth and division. Using a stochastic framework for prediction of the multimeric state of a protein as a function of these processes and the free energy associated with interface-interface binding, we demonstrate agreement with a wide class of proteins using *E. coli* proteome data. As such, this platform, which links protein quaternary structure with biochemical rates governing gene expression, protein association/dissociation and cell growth and division can be extended to evolutionary models for the emergence and diversification of multimers. While it is tempting to think of multimerization as adaptive, the diversity of multimeric states raises the question of its functional role and impact on fitness. As a force driving selection, we consider the possible increase in enzymatic activity of proteins arising strictly as a consequence of interface-interface binding – namely, enhanced stability to degradation, substrate binding affinity, or catalytic rate of multimers with respect to monomers without invoking further conformational changes, as in allostery. For fixed cost of protein production, we find a benefit conferred by multimers that is dependent on context and can therefore become different in diverging lineages.

PACS numbers:

Many of the important features of cellular architecture and function involve the action and interaction of proteins. As such, a comprehensive understanding of the evolution of cellular life necessitates an understanding of the mechanisms by which proteins evolve. A majority of proteins function not as isolated units, but as members of higher-order structures composed of two or more subunits [1]. Many of these complexes are symmetrical and composed of multiple proteins, which are either identical or encoded by the same genetic locus, with each subunit retaining the same catalytic function. These homooligomeric protein assemblies, common in the cellular repertoire, constitute 35% of the bacterial proteome [2]. A survey of the biological protein assemblies from the Protein Data Bank [3] indicates that of the approximately 72,802 proteins examined, $\sim 53\%$ are homooligomers (ranging from dimers to heptamers) [4]. Understanding the evolutionary factors driving homooligomerization is an important goal in the study of protein complexes. While it is tempting to think of the formation of protein complexes as adaptive, this may not be universally the case. It has been observed that in many cases the number of subunits involved in homooligomeric complexes varies among homologues across lineages, with there being no apparent correlation between the size of the complex and the complexity of the organism [5]. This raises the question of the functional role of multimerization and possible impact on fitness.

In a growing cell, the rates of production, degradation, and dilution determine the concentrations of various molecular players, and in turn, the rates of the biochemical reactions in which they take part. The formation of an oligomeric protein requires that individual subunits encounter each other in a crowded cellular environment. Accordingly, oligomer formation depends on the cellular rate constants related to gene expression, multimerization, degradation, and cell growth. Given the stochastic nature of these processes, the number of protein subunits in a cell can fluctuate significantly across a population, even in a homogeneous environment, with implications for cellular functions [6, 7].

In bacterial population genetics, cell growth and division are critical determinants of fitness. At the single-cell level, the growth rate reflects the multidimensional cellular physiological state, which in turns depends on the cell's macromolecular composition. Recent studies support a coarse-grained approach whereby fitness effects of proteins can be mapped to variation in biophysical properties rather than to sequences of mutated proteins, significantly reducing the dimensionality of the genotype-to-phenotype mapping [8–10]. The stochastic nature of protein numbers and resulting effect of their activity on fitness can lead to cell-to-cell variations in growth rate in a genetically identical population of microbes, and this variability can impact the population level fitness [11–13].

In this work, we use a stochastic framework to study the state of homooligomeric proteins in growing and dividing cells and the possible impact of multimerization on fitness. This article is organized as follows. In Section I, we describe the biochemical model governing protein production and multimerization coupled to a realistic model of cell growth and division. The prediction of

[*] Submitted to *Physical Review E*

the multimeric state is shown to agree well with a wide class of homooligomers from the *E. coli* proteome data. While we find that treating cell growth and division as an effective loss term correctly captures the average numbers of monomers and multimers, the stochastic nature of the underlying processes can lead to large variation in protein numbers. Given that the activity of proteins functioning as monomers versus multimers could be different, this variability ultimately impacts cellular fitness and whether or not an emerging multimer is fixed.

In Section II we use this framework to investigate the evolutionary forces driving changes in the multimeric state of a protein across lineages. To highlight this diversity of multimeric states, for the class of small molecule metabolic enzymes of which approximately three hundred are present in both *E. coli* and *S. cerevisiae*, we summarize data across all species to show that most of these shared enzymes exist in both monomeric and multimeric forms. We examine the possible functional benefit to the cell in switching from monomers to multimers while keeping total protein production constant. We consider consequences of multimerization on protein enzymatic activity that result directly from interface-interface binding – namely, increased protein stability to degradation, enhanced ligand binding and catalytic rate. As such, we do not invoke additional conformational changes, such as in allostery, which may require subsequent mutations beyond those affecting only protein-protein interactions and leading to the formation of a novel multimeric interface.

We show that the selective advantage conferred by multimerization resulting from increase in enzymatic activity at fixed cost of protein production depends on context. Specifically, we consider a parsimonious set of parameters governing this fitness benefit (protein abundance, relative stability of multimers to degradation, substrate binding affinity, catalytic rate and concentration of the substrate on which the protein acts) which can become different in diverging lineages. We discuss how these results can account for the variability in the multimeric states of homologous proteins. We conclude with future computational and experimental extensions of this work.

I. GENE EXPRESSION AND MULTIMERIZATION IN THE PRESENCE OF CELL GROWTH AND DIVISION

A. Dimerization

Gene expression and multimerization is described using a four-stage model, depicted schematically in Fig. 1, adapted from that used by Shahrezaei and Swain [14]a. The first stage involves the promoter of the gene of interest, which can transition between inactive and active states, with transcription taking place only while the promoter is active. Activation is followed by transcription and translation, both modeled as first-order chemical reactions defined by a characteristic rate constant. Gene duplication is not explicitly treated, equivalent to using an effective transcription rate. Degradation of mRNA transcripts and proteins during these stages is modeled similarly. Focusing on dimers, where homo-dimers constitute the majority (41%) of oligomeric proteins [4], the final stage consists of protein-protein interactions (PPI), wherein two monomeric protein subunits can bind reversibly to form a dimer, with degradation rates specified for each form.

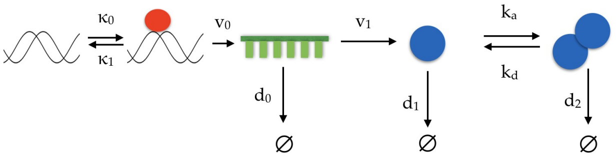

FIG. 1: Biochemical model of gene expression and multimerization. The promoter region of the DNA becomes active (inactive) with probability $\kappa_0(\kappa_1)$ upon binding of the transcription factor (red). Rates of transcription and translation are v_0 and v_1. mRNA (green) and protein (blue) monomers and dimers degrade with probabilities d_0, d_1, and d_2. Dimer association and dissociation rates are given by k_a^D and k_d^D.

Using this framework, we implemented a stochastic algorithm to simulate the protein distributions in a lineage of growing and dividing bacterial cells. The parameters v_0 and v_1 denote the probability per unit time of transcription and translation, respectively; d_0, d_1 and d_2 denote the probability per unit time of degradation of mRNA, protein monomers, and protein dimers, respectively. The association and dissociation of the promoter with the DNA are represented, respectively, by κ_0 and κ_1. The forward and reverse dimerization rates are given by k_a^D and k_d^D, respectively.

Simulations were carried out using parameters summarized in Table I (other parameter sets are given in Table S1). For the dimerization kinetics, the dissociation rate, k_d^D, was set equal to $2 \times 10^{-1} sec^{-1}$, while the association rate, k_a^D, was allowed to vary over values ranging from $k_a^D = 2 \times 10^{-4} M^{-1} sec^{-1}$ to $8 \times 10^{10} M^{-1} sec^{-1}$, with $k_a^D = 10^9 M^{-1} sec^{-1}$ being a typical rate constant [15]. In terms of the free energy of dimer association ΔG, given by

$$\Delta G = RT \ln \frac{K_d}{c_0}, \qquad (1)$$

where $c_0 = 1$ M is the reference concentration, the dimer dissociation constant $K_d = k_d^D / k_a^D$ ranged from 10^{-6} to 10^9 nM, with 10 nM being a value representative of transcription factor homodimers [16].

The increase in cell volume over the course of the cell cycle impacts dimerization by decreasing the concentration of interacting protein subunits, while cell division impacts concentrations via the stochastic partitioning of cellular contents among the two daughter cells. We adopted a model of cell growth and division consistent

Parameter	Definition	Value
v_0	Transcription rate	$5.0 \times 10^{-4} s^{-1}$
v_1	Translation rate	$2.0 \times 10^{-1} s^{-1}$
d_0	mRNA decay rate	$5.0 \times 10^{-3} s^{-1}$
d_1	Monomer decay rate	$5.0 \times 10^{-4} s^{-1}$
d_2	Dimer decay rate	$\leq d_1$
d_3	Tetramer decay rate	$1.3 \times 10^{-4} s^{-1}$
κ_0	Promoter binding rate	$3.0 \times 10^{-4} s^{-1}$
κ_1	Promoter unbinding rate	$1.0 \times 10^{-4} s^{-1}$
k_d^D	Dimer dissociation rate	$2.0 \times 10^{-1} s^{-1}$
k_a^D	Dimer association rate	$(k_d^D/c_0)e^{-\Delta G/RT}$
k_d^T	Tetramer dissociation rate	$2.0 \times 10^{-1} s^{-1}$
k_a^T	Tetramer association rate	$(k_d^T/c_0)e^{-\Delta G/RT}$

TABLE I: Summary of parameter definitions for Section I.

with cell size homeostasis, in which cell volume increases at an exponential rate while the timing of cell division is a stochastic process dependent on both the size and age of the cell [17]. Other theoretical descriptions that go beyond control of cell division based on cell age or size alone, where a constant volume is added at each generation, have also demonstrated agreement with experimental data [18, 19]. The placement of the bacterial contractile ring in cell division has been shown to be tightly controlled, with a standard deviation of 2.9% of cell length in *E. coli* [20], assumed in our simulations. Cellular contents were apportioned randomly according to the binomial distribution, and only one daughter cell was followed at each division. In Supplementary Fig. S4A-C, we plot the resulting distributions of cell length at birth (A) and interdivision time (B), as well as the cell age distribution (C), demonstrating good agreement with previous experimental and simulation results [17–19].

We verify our computational results in the absence of cell growth and division by comparing protein distributions with analytical solutions from [14], as shown in Supplementary Fig. S5. Importantly, we also demonstrate that growth and division affect protein numbers as an additional loss term, showing good agreement between our simulations and analytical distributions in [14] modified according to effective protein decay rates (see below), as shown in the Appendix B and Fig. S6).

For the discrete stochastic chemical kinetics described above, the classical deterministic reaction rate equations, valid in the thermodynamic limit [21], are given by:

$$\frac{dm}{dt} = v_0 P_{on} - d_0 m \quad (2)$$

$$\frac{dM}{dt} = v_1 m - d_1 M - k_f^D M(M-1) + 2k_d^D D \quad (3)$$

$$\frac{dD}{dt} = \frac{k_f^D}{2} M(M-1) - k_d^D D - d_2 D, \quad (4)$$

where m, M, and D, represent the number of mRNA transcripts, and protein monomers and dimers in the cell, respectively; $P_{on} = \kappa_0/(\kappa_0 + \kappa_1)$ is the probability that the promoter for the protein-coding gene is active. Assuming fast mRNA kinetics (relative to the time scale of cell division), the number of mRNA transcripts reaches equilibrium, giving the steady-state mRNA concentration, $\bar{m} = v_0 P_{on}/d_0$, treated as a constant in Eq. 3. In the above, $k_f^D = k_a^D/(N_A V)$, where N_A is Avogadro's number and V is the cell volume.

The population average of the protein concentration, equivalent to averaging the concentration over cells and time in a forward lineage (i.e., according to the population age structure [22], as shown in Fig. S4C), can be obtained as the steady-state solution of the deterministic rate equations. The monomer concentration (in molar units) is given by

$$\bar{c}_M = \frac{1}{2k_a^D}\left[\left(k_f^D - (k_d^D d_1'/d_2') - d_1'\right) + \sqrt{\left(k_f^D - (k_d^D d_1'/d_2') - d_1'\right)^2 + 4k_f^D v_1 \overline{m}\left(1 + k_d^D/d_2'\right)}\right] \quad (5)$$

with $d_1' = d_1 + \lambda$ and $d_2' = d_2 + \lambda$ as effective degradation rates; λ is the dilution rate due to cell division, accounting for the fact that in addition to the intrinsic protein degradation rate, division confers a half-life to proteins given by $\ln 2/\langle T\rangle$, with $\langle T\rangle$ given by the mean interdivision time (consistent with Supplementary Fig. S2). The dimer concentration is given by

$$\bar{c}_D = \frac{(v_1\overline{m}/V') - d_1'\bar{c}_M}{2d_2'} \quad (6)$$

where $V' = N_A\langle V\rangle$.

In Fig. 2, we show the dependence of average monomer and dimer concentrations on the free energy of dimer association ΔG. The box plots show the distribution of single cell averages, demonstrating the noise inherent in gene expression, multimerization and cell division[1]. In Supplementary Fig. S8, we show these results for other parameter sets, including more stable monomers [23]. As expected, the average protein concentrations show a predominance of monomers (dimers) for low (high) negative values of ΔG.

Importantly, this figure shows that while treating cell growth and division as an effective loss term in the deterministic limit of the stochastic model correctly captures the average numbers of monomers and multimers, the stochastic nature of the underlying processes can lead to large variation in their numbers. Given that the activity of proteins functioning as monomers versus multimers could be different, this non-genetic variability results in

[1] In a lineage of dividing cells, the statistics of daughter cells necessarily exhibit correlations due to their shared heritage. In Supplementary Fig. S7, we show the autocorrelation function of protein numbers, demonstrating cyclical variation due to the cell cycle and decay due to noise.

a wide range of phenotypes. The population level fitness can be different from that of the mean phenotype [11], with implications for whether or not an emerging multimer is fixed, addressed in Section II. In Supplementary Fig. S9, we quantify this variability as a function of transcriptional (v_0/d_1) and translational (v_1/d_0) efficiency (Fig. S9A-B), demonstrating agreement with previous works [24]. We also show that the mean and variance of the protein number distributions follow a quadratic relation, consistent with recent works demonstrating this dependence for statistics of protein numbers obtained from snapshots of populations of bacteria and yeast as well as from temporal dynamics of single cells in a lineage [25, 26] (Fig. S9C-D).

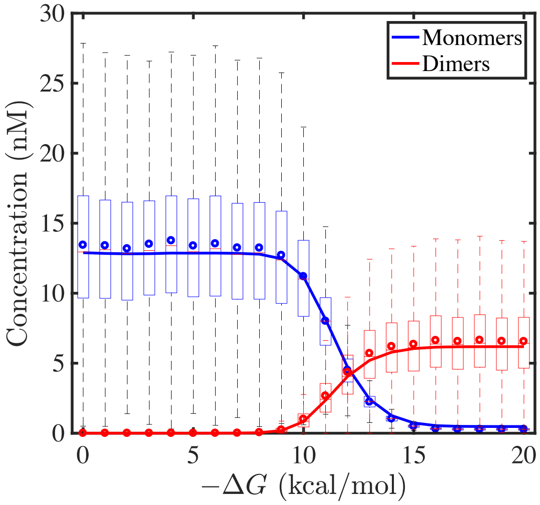

FIG. 2: Average concentrations of monomers (blue/dark) and dimers (red/light) as a function of dimer binding energy, ΔG, for parameter values from Table I. Solid lines show analytical solutions, given by Eqs. 5-6. Boxplots represent data from stochastic simulations for concentrations averaged over individual cell cycles. The central mark is the median, the circle is the mean, and the edges of the box represent 25th and 75th percentiles. The mean concentrations in the box plots agree with those obtained by averaging over the lineage age distribution (Fig. S4C) to within 2.5%. Whiskers extend to the most extreme data points. Data for parameter sets B-D shown in Supplementary Fig. S8.

In all simulations, the decay rate of the dimeric form of the protein is assumed to be lower than that of the monomeric form. It has been argued that larger proteins are more stable due to their extensive internal interactions as well as reduced surface area/core ratio which leads to reduced solvent exposure [27]. However larger proteins are more difficult to maintain, given the greater likelihood of transcript errors in a longer protein sequence, but the same stability advantage can be achieved by oligomerization, resulting in a large protein composed of several short-sequence monomeric subunits [2, 28, 29]. Our results show that with increasing rate of dimerization, the total amount of protein present in the cell increases, independent of the rates of transcription and translation. This is due to the fact that dimerization protects the enzyme from degradation, allowing larger amounts to accumulate in the cell over the course of the cell cycle. As discussed in Section II, this result also has implications for the evolutionary dynamics of the protein interface leading to multimerization, when the fitness of the cell is related to the total enzymatic activity of the protein. Under such circumstances, the transition from a monomeric enzyme to a homomeric form could confer a selective advantage simply by resulting in an increased amount of enzyme present in the cell for fixed cost of protein production, even without any change in the catalytic efficiency of the multimeric form.

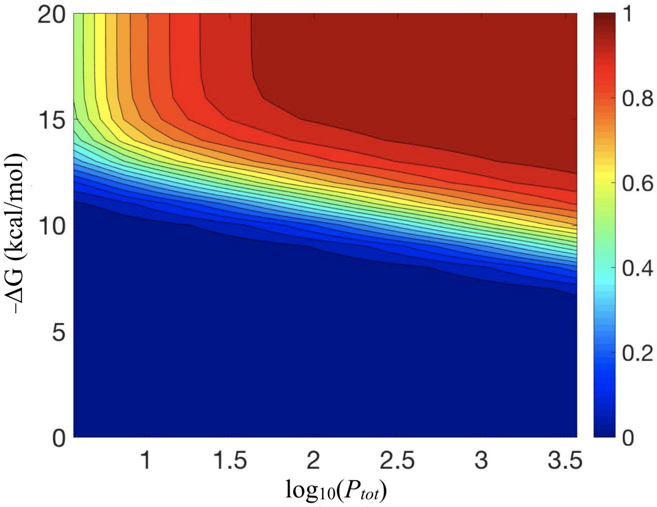

FIG. 3: Dimer fraction as a function of ΔG and total protein number P_{tot}. P_{tot} is adjusted by varying v_1, with all other parameters held fixed at values given in Table I.

In Fig. 3, the impact of the protein dimerization rate (expressed in terms of the interface binding energy, ΔG) and total expression level, P_{tot}, on the fraction of proteins that exist as dimers is shown. For the governing parameters given by Table I, the proteome transitions from a predominantly monomeric state to a predominantly dimeric state at $\Delta G \approx 10\,\text{kcal/mol}$, and this transition will shift to moderately lower values of ΔG for higher expression levels. Additionally, protein localization could promote dimerization at lower ΔG values, though non-functional interactions with other proteins may have the opposite effect, requiring larger ΔG in order to achieve higher specificity (see Table II). Overall, this framework, extended below to include tetramers, allows for prediction of the predominant quaternary state of a protein given the expression level of its coding gene and the strength of multimerization.

B. Tetramerization

Higher-order complexes generally assemble via ordered pathways [30], with monomers assembling into dimers, which then assemble into tetramers, etc. If proteins can be treated as having two independent and non-overlapping interfaces that participate in the formation of oligomers, the biochemical model is easily modified to include tetramerization, characterized by rates of dimer association and dissociation:

$$\frac{dM}{dt} = v_1\bar{m} - d_1 M - k_f^D M\left(M-1\right) + 2k_d^D D \quad (7)$$

$$\frac{dD}{dt} = \frac{k_f^D}{2} M\left(M-1\right) - \left(k_d^D + d_2\right) D - k_f^T D\left(D-1\right) + 2k_d^T T \quad (8)$$

$$\frac{dT}{dt} = \frac{k_f^T}{2} D\left(D-1\right) - \left(k_d^T + d_3\right) T \quad (9)$$

where $k_f^T = k_a^T/\left(N_A V\right)$ and k_d^T are the forward and reverse tetramerization rates, and d_3 is the decay rate of tetramers. Coupled to cell growth and division, the average abundance of monomers ($\bar{M}$), dimers ($\bar{D}$), and tetramers ($\bar{T}$) in the cell can be solved numerically as before, from Eqs. 7-9, as shown in Supplementary Fig. S10.

C. Comparison with *E. coli* proteome data

The output of the model was compared with *E. coli* proteome data from [31] and ΔG values from the PDBePISA database [32] (Table II). A sample of proteins was chosen for which data on cellular abundance and structure were available, and which had a recorded assembly with either 1, 2, or 4 subunits. The recorded assembly of each protein from the PDB was compared with the distribution of monomers, dimers, and tetramers predicted by our model. The model is in favorable agreement with the experimental data, correctly predicting the oligomeric assembly for 91% of the proteins tested.

Some discrepancies do exist between the model and the reported oligomeric state of some enzymes. For instance, *glf* (glucose facilitated diffusion protein) is reported as a dimer in the PDB, however our model (as well as PDBePISA) predicts a monomer. *gpmA* (glycolysis pathway protein) is also reported as a dimer, but predicted to be a monomer by the model. In this case, the assembly appears to be moderated by two chlorine ions and four sulfates giving rise to an overall $\Delta G < -110$ kcal/mol, with the protein-protein interface contributing only -4.9 kcal/mol to the total. *secB* (molecular chaperone involved in protein export) simply fails the model, as the reported ΔG is not strong enough to promote the formation of dimers in our scheme.

These disagreements between the model prediction and the given assembly could arise for a number of reasons. The existence of post-translational modifications (as in the case of *gpmA*) or ligand-induced conformational changes in the protein may impact oligomerization in a way that is not captured by the model. Protein localization may lead to more oligomerization than is predicted by effectively increasing the concentration, and promiscuous PPI may interfere with functional interactions. It is interesting to note that many dimers have free energies well beyond the values at which dimers start to outnumber monomers in the model, suggesting the presence of a pressure toward stronger interfaces. This could potentially reflect a selective benefit conferred due to minimizing non-functional PPI that place limits on gene expression and protein diversity [33]. Strengthening an interface in a multimeric enzyme beyond the extent necessary to ensure dimerization can also stabilize the active site, thus improving substrate binding to the protein [34].

A useful representation of these results is the composition of the proteome as a function of the two (potential) interface free energies. If it is assumed that tetramers assemble via a single pathway (so that there exists only one type of intermediate dimer complex), the characteristics of the proteome can be predicted well by the model, as described above. It has been shown that protein complexes tend to assemble via specific ordered pathways with the larger interface forming first [35], though this does not always correspond to the interface with the larger free energy gain [36].

Supplementary Fig. S11 shows the fraction of proteins in each of the three multimeric states (monomer, dimer, and tetramer) as a function of the free energy of the two interfaces, ΔG_1 and ΔG_2. An interesting finding is that for most values of ΔG_1 and ΔG_2, the proteins exist almost entirely in one particular quaternary structure, rather than in an equilibrium consisting of significant amounts of all three types of assembly. Deviations from this result certainly exist, however, with one possible cause being phenotypic noise, which will be higher for proteins with a higher translation rate (Fig. S9).

In summary, the results of Section I show that the multimeric state of a protein can be predicted from the relevant biochemical parameters. The interface energy marking the transition from monomeric to multimeric form varies inversely with abundance, suggesting a potential tradeoff wherein a decrease in expression necessitates a strengthening of the interface. In Section II, we apply these results to the evolution of multimers from monomeric subunits to address whether changes in the underlying biochemical parameters affecting the multimeric state will impact cellular fitness. Once a mutation occurs that alters protein abundance (through change in transcription or translation rate) or the interface binding energy, this mutation may become fixed in a population if the multimer is sufficiently beneficial. Stochastic noise introduces significant variability in the numbers of monomers and multimers, which in turn affects the putative strength of selection for multimers. We construct a measure of the fitness benefit for emerging multimers and investigate its dependence on underlying governing parameters to understand the observed variation in the state of a particular protein from one species to another.

II. FUNCTIONAL ROLE OF MULTIMERIZATION

Across the tree of life, homologous proteins with identical function can form complexes with differing number of subunits in different organisms, with there being no apparent correlation between organismal complexity and number of subunits [5]. Examples include several glycolytic enzymes (such as hexokinase, phosphofructokinase, and phosphoglucomutase) that form multimers in some species but remain as monomers in others [5], as well as other highly conserved enzymes involved in nucleotide (such as dihydrofolate reductase, adenosime deaminase, and guanylate kinase) and lipid metabolism (such as acetyl-CoA carboxylase, hormonse-sensitive lipase, and dodecenoyl-CoA isomerase), and peptidoglycan biosynthesis (such as peptidoglycan glycosyltransferase and glutamate racemase).

To further highlight this diversity of multimeric states, in Table S2 we examine the class of small molecule metabolic enzymes, of which 271 are present in both *E. coli* and *S. cerevisiae* [37]. We find, using data across all species from the PDB and BRENDA databases, that as many as 65% of these shared enzymes exist in both monomeric and multimeric forms across the tree of life, with some orthologous enzymes ranging from one to more than ten subunits. While it has been suggested that protein-protein interfaces evolve to optimize association with respect to the protein's function [38], these data bring under question whether multimerization presents the organism with a functional advantage in turn affecting fitness, which we address in this section.

A ubiquitous property of multimeric proteins is allostery, where interactions between subunits result in cooperativity in the function of the protein. In many cases the role of cooperativity is to confer enhanced regulation by allowing sensitive response to change in concentration of an internal or external signal in many cellular signal transducing systems as well as enzymes (where most metabolic enzymes are allosteric [39]. Despite its importance in many biological processes, the structural principles underlying cooperativity in multimeric proteins and the evolutionary origins of allosteric communication between subunits are not fully understood [40].

Recent work has shown that a small fraction of amino acids comprise a spatially distributed but structurally contiguous sub-network within the tertiary structure. These co-evolving networks, dubbed "sectors", have been shown in several protein families to be associated with allostery, where the spatial distribution of sectors effectively "wires" the protein's active site to multiple distant surface positions [41]. Hence, a mutation that creates a novel protein-protein interface would not necessarily immediately bring with it ancillary effects consistent with allostery. As a concrete example, the enzyme DHDPS (dihydrodipicolinate synthase) which initiates lysine biosynthesis, is primarily a homotetramer in all species, but the architecture of the tetramer differs across kingdoms, functioning allosterically in some but not others [42, 43]. Rather, it is likely that the structural basis of a protein's cooperative function relies on subsequent mutations beyond the formation of a new PPI, and therefore would not be manifest immediately upon multimerization. Therefore, in addressing the functional role of multimerization, we assume no allosteric effects in the activity of proteins and consider only direct consequences of interface-interface binding, as discussed below. We are nonetheless able to show that even in the absence of allostery, there can be a quantitative benefit to multimerization that is context-dependent, and subsequent mutations conferring allosteric advantages could further solidify the advantage of the multimer, but might not occur right away or in every case.

A. Fitness Advantage of Multimers

We connect multimerization to cellular fitness through enzymatic activity assuming Michaelis-Menten kinetics, with protein abundances in growing and dividing cells given in Section I. We focus on enzymes in which each subunit is catalytically active. For enzymes which require a minimal multimeric form to be active, or enzymes in which the active site forms as a result of a dimeric interface, this analysis would be used to compare, for example, dimers and tetramers, rather than monomers and dimers. By examining the impact on cellular fitness under different conditions determined by relevant governing parameters, we make connection with the diversity of multimeric states of a protein in different species.

Focusing on homodimers, which constitute the majority of multimeric proteins, we construct a dedimensionalized measure of activity, $\hat{\alpha}$ (in units of $k_{cat\,M} \cdot K_m^M$), which can be written as:

$$\hat{\alpha}(\hat{c}) = \hat{c}_M \frac{\hat{c}}{1+\hat{c}} + 2\hat{c}_D\, r_{cat}\, \frac{r\hat{c}}{1+r\hat{c}}, \tag{10}$$

where $r = K_m^M / K_m^D$ with K_m^M and K_m^D given by the Michaelis-Menten constants for monomers and dimers, respectively, such that $K_m^{M,D} = (k_{-1}^{M,D} + k_{cat}^{M,D})/k_1^{M,D}$ and $k_{\pm 1}^{M,D}$ are the respective binding/unbinding rates; $\hat{c}_M = c_M / K_m^M$ and $\hat{c}_D = c_D / K_m^M$ are the monomer and dimer concentrations, respectively, while $\hat{c} = c/K_m^M$ is the free (unbound) substrate concentration; $k_{cat\,M}$ and $k_{cat\,D}$ are the catalytic rates of the subunits in monomeric and dimeric forms; and $r_{cat} = k_{cat\,D}/k_{cat\,M}$, is the ratio of catalytic rate of the dimeric form to the monomeric form of the protein. We note that in Eq. 10, the concentration of free substrate, $\hat{c}$, will depend on the total substrate concentration, as well as on the concentrations of bound monomers and dimers.

Manufacturing proteins consumes cellular resources, both from the standpoint of using energy and nutrients as building blocks and also by occupying common cellular machineries such as ribosomes, polymerases, or chaperones [44]. In turn, a given protein contributes to the cell's fitness through its function in the cellular repertoire. To investigate the possible impact of multimerization on cellular fitness, we consider the benefit to and cost incurred by the cell in producing multimers, with benefit measured in terms of total activity and cost in terms of total protein production. Specifically, we consider two scenarios, with and without multimerization, in which the total number of proteins produced, P_{tot}, and hence the cost associated with their production, is the same. However, the total activity of functional proteins, and therefore the benefit associated with their function, may be different in the two cases. This is a relevant basis for comparison, as a mutation that creates a dimeric interface is unlikely to alter the amount of protein produced by the cell.

Defining the metric ϕ as the effective energetic advantage of multimerization relative to the total proteome energy budget of the cell:

$$\phi = \frac{\beta\left(\hat{\alpha}_{tot} - \hat{\alpha}_{tot,0}\right)}{E_{tot}} = \frac{\beta\Delta\hat{\alpha}_{tot}}{E_{tot}}, \tag{11}$$

where $\hat{\alpha}_{tot}(= V\hat{\alpha})$ represents the total enzymatic activity per cell with multimerization; $\hat{\alpha}_{tot,0}$ represents the corresponding activity in a hypothetical "reference cell" in which there is no multimerization. The parameter β characterizes the benefit of the protein of interest (given in units of ATP produced, or equivalent), and E_{tot} is the total energy budget of the cell in units of ATP hydrolyses, approximately 27×10^9 for *E. coli* [45]. Here, we have assumed a simple benefit function that is linear in protein activity, and cost that is linear in protein production given by P_{tot}, assumed to be the same with and without multimerization, where in *E. coli*, P_{tot}, ranges from < 1 to 8000 [31]. Other related approaches to assigning cost and benefit for specific proteins [9, 46–48], and metabolic networks [49] have been formulated.

In bacterial populations, the growth rate equates directly with fitness as faster growing cells outgrow competitors. With the economy of protein production ultimately linked to cell growth and division, we can relate the duration of the cell cycle, t_D, to a cellular (or microscopic fitness function [47]), $w = 1 + s$, as

$$t_D = t_{D,0}/w = t_{D,0}/\left(1 + \gamma\phi\right) \tag{12}$$

where the selection coefficient, $s = \gamma\phi$, is assumed to be proportional to the energetic benefit of dimerization.[2] We consider $w(s = 0) = w_0 = 1$ to be the fitness in the absence of multimerization, with $t_{D,0}$ giving the associated cell cycle duration. The relative increase in growth rate due to multimerization is therefore

$$\frac{\Delta\tilde{\lambda}}{\tilde{\lambda}_0} \approx \gamma\phi = \gamma\beta\frac{\Delta\hat{\alpha}_{tot}}{E_{tot}}, \tag{13}$$

where $\tilde{\lambda}_0 = 1/t_{D,0}$ is the cellular growth rate in the absence of multimers. Variation in growth rate is thus tied to variation in activity, consistent with experimental results showing that fluctuations in expression of enzymatic proteins can lead to fluctuations in cellular growth rates, and vice versa [50].

Noting the proportionality of the growth rate advantage to the product $\gamma\beta$, we therefore explore $\Delta\hat{\alpha}_{tot}/E_{tot}$ as a function of governing variables, as described below. We can interpret this expression, rewritten as $(\Delta\hat{\alpha}_{tot}/\nu P_{tot})\,(\nu P_{tot}/E_{tot})$, in terms of the increase in total activity per energy spent on manufacturing protein, times the relative fraction of the cellular energy budget consumed by the protein of interest, where the constant ν represents the metabolic cost (per protein) to the cell of manufacturing proteins (estimates for ν yield values on the order of 5000 ATP per average protein [45]). This results in a measure that is proportional to total activity rather than activity per protein, which is consistent with results showing that the fitness benefit of an increase in catalytic efficiency can be offset by a decrease in protein abundance [8].

In the absence of allosteric effects, there are three mechanisms by which dimers can have increased activity with respect to monomers, leading to a growth rate advantage. The first is through a decrease in the dimer decay rate relative to that of monomers ($d_2 < d_1$), which may occur if a dimer is more stable to degradation than its individual subunits, or if the monomer is targeted by a highly specific protease at a cleavage site that is covered by the dimer interface. A more stable dimer will result in a higher overall protein level in the cell, and thus higher activity.

The second is enhanced enzymatic activity as a result of increase in the rate of substrate binding, $k_1^D > k_1^M$. How can this be achieved without requiring conformational change of the multimer or its monomeric subunits

[2] We expect the energetic advantage of a given protein to be a small fraction of the total proteome energy budget, $\phi \ll 1$. Most generally, the selection coefficient will be a function of ϕ, $s = g(\phi)$, relating the cell cycle duration to the energetic advantage of producing multimers, and its form will depend in detail on the mechanism(s) by which a given protein impacts the cell cycle. However, we can approximate $g(\phi) = g'(0)\,\phi + \mathcal{O}(\phi^2) \approx \gamma\phi$, where $\gamma > 0$ is a constant that measures how strongly the growth rate varies in response to the production of a given protein. This formulation is consistent with previous works on the impact of the cost and benefit of protein production on growth rate, specifically for Lac proteins [46] and MetE [48] in *E. coli*, where the underlying parameters are determined from fits to experimental data.

as in cooperativity? It has been shown that the formation [43, 51] or strengthening [34] of a protein-protein interface can reduce dynamic fluctuations near the binding site, thus increasing its specificity. Furthermore, as discussed in Appendix A, if binding of substrate to its target site on the protein (monomer or multimer) is governed by a two-stage process – whereby three-dimensional diffusion of substrate in the cytosol brings it within interaction distance of the protein, followed by adsorption to the protein surface and subsequent diffusion in two dimensions to the binding site – it is possible for the capture rate per binding site to be greater for multimers than for monomers [52, 53]. This results in a lower K_m for dimers, such that $r = K_m^M/K_m^D > 1$, and increased dimer activity.

A third potential mechanism to increase dimer activity could come from an increased catalytic rate, $r_{cat} > 1$. Indeed, stabilization of the active site upon interface-interface binding can not only impact the substrate binding residues, leading to enhanced substate binding as discussed above, but also the catalytic residues involved in modifying the substrate. As an example, this has been postulated for the tetrameric enzyme DHDPS [43] (which has been established to not be allosteric in *E. coli*), where it is observed to have a higher catalytic rate in tetrameric than dimeric form.

Below, we extend the results of Section I to investigate the possible fitness advantage of multimerization as a function of governing parameters. According to Eq. 10, the total activity, $\hat{\alpha}_{tot}$, depends on the monomer and multimer concentrations. Both concentrations increase with increasing total protein production, P_{tot}, while the strength of multimerization, determined by ΔG, and ratio of decay rates, d_1/d_2, set their relative concentrations for given protein production. We also note the dependence of enzymatic activity on the concentration of substrate, $\hat{c}$, which is in turn controlled by internal and/or environmental conditions. Indeed, protein production is tightly coordinated with external conditions and intracellular demands, and many enzymatic proteins are saturated by their substrates because less protein is required to achieve the same rate of product formation. By ensuring that proteins are produced at a needed level, this regulation, which we do not incorporate into our analysis in this work, serves to minimize the overall cost of protein production. Finally, the ratio of Michaelis-Menten constants, $r \geq 1$, and catalytic rates, $r_{cat} \geq 1$, govern the enzymatic kinetics. We explore the fitness advantage as a function of parameters determining the relative concentration of multimers, d_1/d_2, P_{tot}, and ΔG, as well as those governing enzymatic kinetics, namely r, r_{cat}, and $\hat{c}$.

B. Connection to Diversity of Multimeric States

To understand the potential benefit of multimerization, we examine various mechanisms of increase in activity separately. We first consider the case where enhanced dimer stability to degradation is the only source of any fitness benefit ($r = 1$, $r_{cat} = 1$ and $d_2 < d_1$) as shown in Fig. 4A. When there is significant dimerization ($\Delta G < -10$ kcal/mol), the fitness benefit increases with substrate concentration, $\hat{c}$, saturating to a maximum when the proteins are saturated. This maximum fitness benefit increases with increasing expression level, P_{tot}. Similar trends result when only the increased catalytic rate confers benefit (Fig. 4B).

On the other hand, if enhanced substrate binding to dimers is the cause of increased activity ($d_2 = d_1$, $r_{cat} = 1$ and $r > 1$), as in Fig. 4C, the benefit peaks at an optimum value of $\hat{c}$, then vanishes when $\hat{c}$ becomes high enough to saturate monomers in the reference cell. In this case, the advantage conferred by multimerization is highly dependent on substrate concentration; indeed, $\hat{c}$ spans several orders of magnitude for metabolites involved in core metabolism in *E. coli*, with a majority having values between 10^{-1} and 10^{1} [54] as shown in this plot. Therefore, a novel dimer may or may not confer a selective advantage to reach fixation based on this effect alone. Fig. 4D shows the combination of the effects in (B) and (C), resulting in a higher benefit that peaks at lower values of $\hat{c}$. While in general these effects may appear in combination, we make a parsimonious choice of the functional consequences of multimerization in subsequent plots, assuming no enhanced catalytic activity ($r_{cat} = 1$) which in some cases may require additional conformational changes.

For fixed $\hat{c}$, the benefit will increase with increasing r and P_{tot} (results not shown). For low P_{tot}, both $\hat{\alpha}_{tot}$ and $\hat{\alpha}_{tot,0}$ will be negligible, and the benefit will be small; for large P_{tot}, all substrate will be bound in both scenarios (with and without multimerization), and the benefit will again be small. Thus, there will be an optimal value of P_{tot} that maximizes the benefit, which occurs when all substrate molecules are bound to protein in the presence of multimerization, but not yet in the reference cell. This optimum P_{tot} (which for the parameters of Table I lies outside the typical physical range of protein numbers shown in Fig. 4) shifts to higher values with increasing $\hat{c}$, as more protein is needed in order to bind all of the substrate; the peak magnitude of the fitness advantage increases in this case.

In Supplementary Fig. S12, we use data from BRENDA to plot K_m values for core metabolic enzymes in *E. coli*, whose absolute metabolite concentrations were measured in recent work [54]. We note that most have K_m^M values between 10^{-6} M and 10^{-2} M, with a median value of 1.7×10^{-4} M, which we use in our results. Since approximately 83% have $\hat{c} > 1$ [54], this suggests that for this class of proteins, enhanced substrate binding is not a primary driver of increased dimer activity. Some additional advantage such as higher dimer stability to degradation or increased catalytic rate would be needed to give a dimer-producing phenotype a significant selective advantage.

These results account for the apparent contradiction of

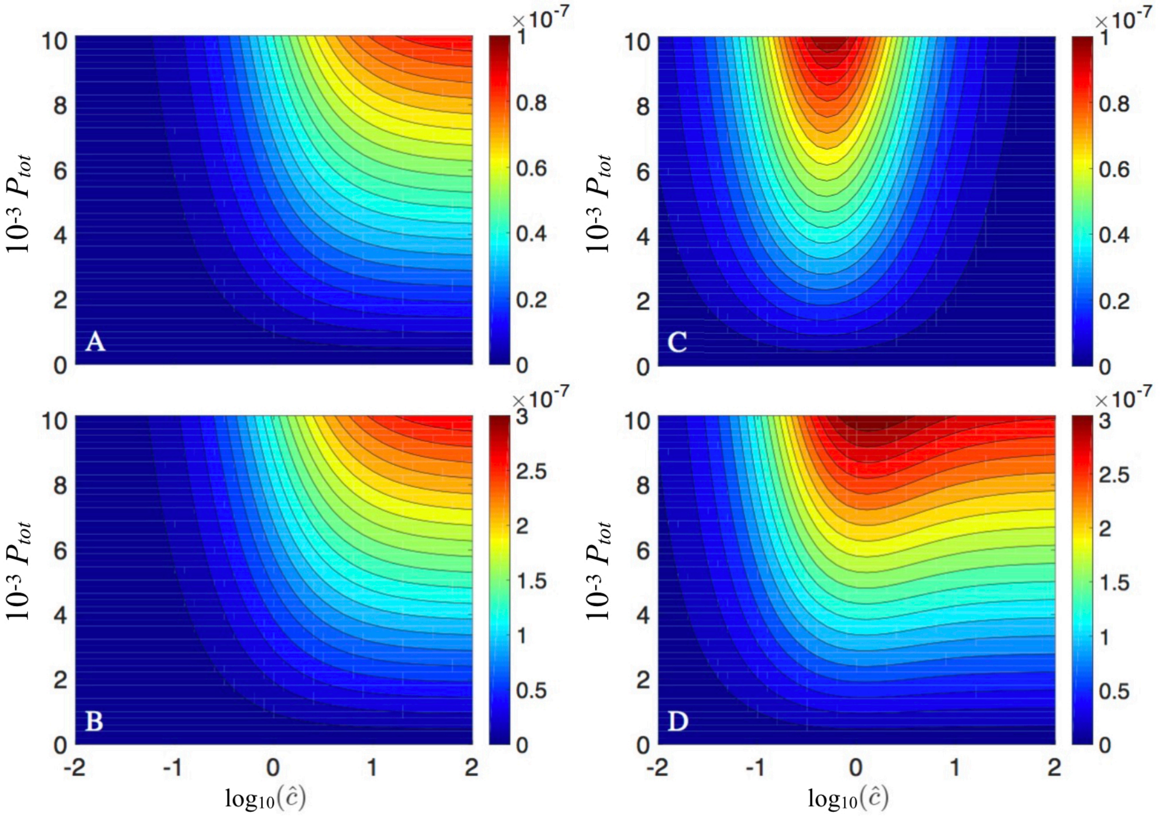

FIG. 4: Contour plots showing increase in protein activity normalized to total cellular protein budget, $\Delta\hat{\alpha}/E_{tot}$, as a function of total protein production. The first three plots show scenarios in which there is only one mechanism of increased activity: (A) $d_2 = d_1/2$, $r = 1$, $r_{cat} = 1$: As a function of $\hat{c}$, the benefit is negligible for small values when little substrate is bound, and saturates to a maximum when all substrate is bound to protein. (B) $d_1 = d_2$, $r = 1$, $r_{cat} = 2$: Trends are the same as in (A). (C) $d_1 = d_2$, $r = 5$, $r_{cat} = 1$: The benefit derives from enhanced substrate binding to the dimer; it peaks as a function of $\hat{c}$ and goes to zero at high concentrations when all dimers and monomers are bound. (D) $r = 5$ and $r_{cat} = 2$: The trend is similar to that in (A) and (B), except that the benefit peaks at a lower substrate concentration due to the enhanced binding. $\Delta G = -20$ kcal/mol and $K_m^M = 10^{-4}$ M in all plots.

the putative advantage conferred by multimerization and the persistence or even reemergence of monomery across lineages, where dimerization (more generally, multimerization) of a particular enzyme may be highly beneficial in one species, but less so in another, perhaps even closely related, species. Additionally, they demonstrate how mutation bias affecting transcription or translation rates, or substrate binding, may push an enzyme from a regime in which dimers are clearly advantageous to one in which the increase in activity is negligible. This can result in a pressure to decrease the magnitude of ΔG, since the formation of a dimer interface involves an increase in surface hydrophobicity, which presents a risk of promiscuous interactions; again, this risk may also vary between species. Differences in protein abundance resulting from interspecies differences in expression level may also make multimerization more likely to be beneficial in some species than in others. These and other factors may mitigate any adaptive benefit of multimerization and account for the observed variability in the multimeric states of homologous proteins.

In Figs. 5 and 6, we consider the variability in the fitness advantage as a result of the stochastic dynamics of multimers in growing and dividing cells. The population average of the fitness benefit from simulations agrees well with the analytical result obtained using the protein concentrations given by Eqs. 5-6. In Fig. 5, we show the fitness benefit as a function of ΔG for $r = 1, d_1/d_2 = 2$ and $\hat{c} = 1, 10$. We note that while the mean population response shows a clear difference between the monomer and dimer fitness advantage, there is significant variability within a population whereby fluctuations in monomer and dimer numbers render this advantage ambiguous. This ambiguity, coupled with the risk of promiscuous interactions, may impact the long-term advantage of a mutation that lowers ΔG. This variability will depend on various biochemical parameters, such as transcriptional and translational efficiency, as shown in Supplementary Fig. S9.

In Fig. 6, we note that when the dimer fitness advan-

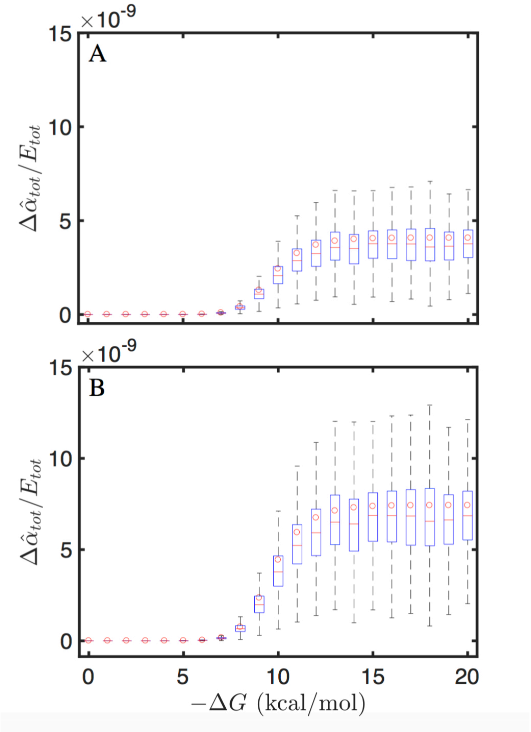

FIG. 5: Increase in protein activity as a function of ΔG for (A) $\hat{c} = 1$ and (B) $\hat{c} = 10$, with r=1, $d_1 = 2d_2$, and $P_{tot} \approx 625$ in both plots. Circles denote analytical results and the box plot is from simulations. When derived from increased dimer stability to degradation, this fitness benefit at fixed cost increases as the relative abundance of dimers increases (with increasing interface binding energy) and improves with increasing enzyme saturation (increasing $\hat{c}$).

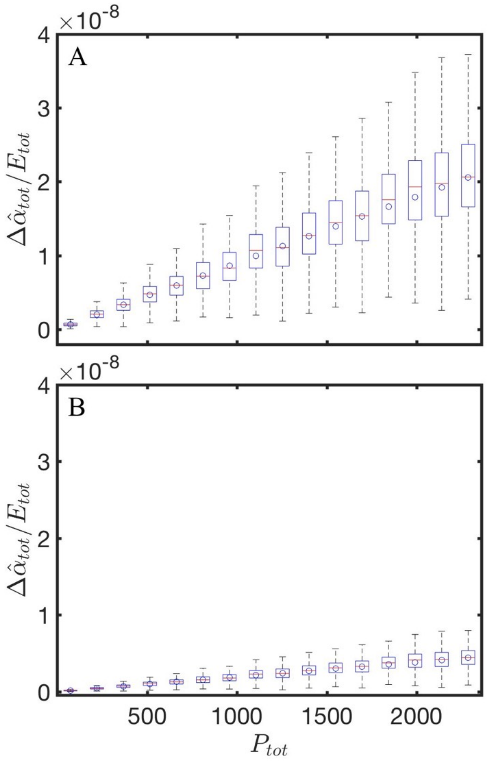

FIG. 6: Increase in protein activity as a function of P_{tot} for $d_1 = d_2$, r=5, $\Delta G = -20$ kcal/mol, and (A) $\hat{c} = 1$ and (B) $\hat{c} = 10$. Circles denote analytical results and the box plot is from simulations. In the absence of relative stability of dimers to degradation, at higher substrate concentrations such that both monomers and dimers are saturated, the fitness benefit is diminished.

tage is due to enhanced substrate binding ($d_1 = d_2, r = 5$), it decreases significantly when $\hat{c} = 10$, consistent with Fig. 4B: When the enzyme is saturated with substrate in both dimeric and monomeric forms, any advantage due to enhanced binding is diminished. On the other hand, when the fitness advantage is due to enhanced stability of dimers to degradation relative to monomers, Fig. 5 shows that it is greater at higher substrate concentrations given higher relative abundance of dimers. These results suggest that the precise dependence of the fitness advantage on $\hat{c}$ will depend on the nature of the mechanisms conferring the advantage, as well as on the specific biochemical parameters.

In these plots, it is clear that the non-genetic variation arising from stochastic dynamics of multimers results in a wide range of phenotypes, with the fitness advantage approaching zero in some individuals. This variability can have profound implications for the evolutionary fate of a newly formed multimeric interface. Even after a new PPI has become widespread, the population-level performance can be greater or less than that suggested by the mean phenotype, depending on the nature (convexity or concavity, respectively) of the function relating phenotype to fitness, as described by Jensen's inequality [11]. This function will depend on the specific protein of interest, and may vary across species for a given protein, with implications for the evolutionary diversity in multimeric state observed for many proteins.

III. DISCUSSION

Homooligomeric proteins form an important class of protein complexes in the cellular repertoire. Their formation on cellular time scales requires successful interactions between subunits: Here, we build on previous work [14] to develop a stochastic model of gene expression and multimerization in growing and dividing cells, using an experimentally realistic model of cell division. Cell population averages agree well with the steady state solution of the deterministic limit of the stochastic model, with the effect of cell division on protein numbers incorporated as an additional, effective loss term. Furthermore, we show that when this effective loss term is included, previous analytical results for the protein distribution absent cell division [14] agree with distributions from our simulations. The multimeric states predicted by the model, obtained using biochemical parameter values from the experimental literature while adjusting the translation rate to achieve the experimentally determined total protein number, agree well with a wide class of homooligomers from the *E. coli* proteome data.

Recent work has emphasized a coarse-grained approach to genotype-to-phenotype mapping, where fitness effects of mutations at the sequence level of a protein are projected onto a small number of axes representing its biophysical and biochemical properties [55], such as stability, substrate binding affinity and catalytic activity [9]. Several arguments have been put forward to explain the adaptive benefit of multimers, including increased encounter rates between enzyme and substrate (where diffusion-limited rates are proportional to the effective radius of the enzyme [56]), stabilization of catalytic sites [43] or the formation of new active sites [2], protein activation [1], protection against denaturation [28], protection against aggregation in thermophilic organisms [55], and allosteric regulation [57]. Despite these potential advantages, a number of counterarguments could be made disputing the adaptive benefit of multimerization. First, there exist multimers that appear to function no more efficiently in their host organism than monomeric forms of the same enzyme in other organisms [58]. Second, the emergence of a new protein-protein interface can lead to promiscuous PPI. This can lead to aggregation, which has been implicated in a number of diseases [59]. It is conceivable that mutations that create a new multimeric interface will incur a cost due to increased surface hydrophobicity, which can lead to promiscuous interactions, yet will become fixed as long as the novel multimer is sufficiently beneficial. Otherwise, these mutations will have an overall negative effect on fitness and be eliminated by selection.

Here, we considered the role of total enzymatic activity in driving multimerization. We show that for a fixed cost of protein production, multimerization affects enzymatic activity in a context-dependent manner. Focusing on homodimers, we find that for given underlying parameters governing protein production (namely rates of transcription, translation and degradation) this fitness advantage is necessarily dependent on (i) the relative concentration of dimers, set by the free energy of dimerization, and relative stability of dimers to degradation, (ii) the binding affinity for substrate and catalytic rate, which may be higher for dimers than for monomers as a direct consequence of stabilization of the active site upon interface-interface binding, even without allostery, as discussed in Sections II and IV, and (iii) the concentration of substrate. This context-dependence of the fitness advantage is consistent with the diversity of multimeric states of some highly conserved enzymes, as the benefit can become very different in diverging lineages. It can also impact the fixation of mutations that change the multimeric state; if a dimer interface forms in a cell in which monomers and dimers lead to nearly identical fitness, the fixation of the mutation will depend ultimately on random genetic drift. The same factors could also lead to a reversion of a dimeric enzyme to monomeric form, or alternatively, progression to higher order multimers. Indeed, as a relevant illustration of the diversity of multimeric states, our survey of PDB and Brenda databases of small molecule metabolic enzymes shared between *E. coli* and *S. cerevisiae* revealed that 65% exist in both monomeric and multimeric forms across the tree of life, with some orthologous enzymes ranging from one to more than ten subunits.

In considering the possible fitness advantage of multimers resulting from enhanced total activity, we note that most generally, cellular processes do not work in isolation, but rather the function of a given protein necessarily interacts with the rest of the cell. Recent work coupling gene expression to cellular and population growth has considered fundamental tradeoffs, such as limitations in levels of cellular energy, free ribosomes, and proteins, that affect the evolution of the expression level of certain proteins [60]. Future extensions of the present work on the selective benefits of multimerization on fitness would similarly consider mechanistic links between trade-offs, gene expression and growth.

As expected, our results show that the noise inherent in protein production, multimerization and cell division introduces variability in the fitness benefit at the population level. While the present work does not explicitly couple growth related parameters to the physiological state of the cell – here, given in terms of the total activity of a given protein of interest – within a clonal microbial population, this noise in protein levels and their associated activity will introduce variability in growth rate among individual cells.

Recent experimental [13] and theoretical [12] works have addressed the effect of this heterogeneity on the population growth rate, demonstrating that it can lead to faster or slower population growth than the single cell mean. In the context of the diversity of multimeric states, these results suggest that the population level fitness can be affected on the one hand by change in the mean single-cell growth rate resulting from a possible functional advantage of multimers, and on the other hand by noise-driven growth rate effects. The latter effect can be positive or negative, and has been shown theoretically to depend on the growth rate variability and the strength of the correlation between mother and daughter division times [12]. This correlation may depend on several factors, including the expression level of the protein and stability to degradation (which in turn affects how many proteins are inherited as opposed to produced by each cell), and the degree to which the protein of interest affects growth rate. The present work – whereby stochastic protein expression, multimerization and activity occur in growing and dividing cells using a realistic model of cell growth and division, necessarily with correlations between mother-daughter growth rates to account for cell size homeostasis – provides a computational framework for future extensions that couple the single cell growth rate to the multimeric state, allowing systematic investigation of these effects on the population level fitness.

On the experimental front, our work suggests several extensions. Site-directed mutagenesis aimed at alterring targeted sites in model homooligomeric enzymes with well-characterized biochemistry can in principle allow systematic experimental studies of the possible relation between multimerization, stability to degradation, enzymatic activity, and cellular growth rate. The effects of modifications to catalytic and substrate-binding residues, as well as residues altering the interface binding energy ΔG and disrupting multimerization with [43] or without [61] accompanying change in function can be directly addressed within the framework presented here. Furthermore, recent studies have highlighted the role of “indirect mutations" at residues not at the interface itself in changing the interface binding energy through change in intersubunit geometry [29, 62]. If the distant residues are involved in substrate binding, for example, our framework can be extended to include an associated change in ΔG that would help or hinder multimerization.

Related to this, previous work, primarily on monomeric enzymes, supports the idea that residues in a protein that participate in catalysis are not optimized for folding stability: Indeed, it has been shown that the existence of active sites introduces strain in the protein structure, making the folded state less stable, and conversely, it is possible to stabilize proteins by sacrificing activity [63–65]. Recent work postulated a fitness landscape for a (monomeric) protein based on two biophysical traits given by the folding stability and binding affinity to a target molecule [66]. The evolutionary dynamics in this fitness landscape simulated under the sequential model demonstrated that as a result of the coupling between folding and binding – where only folded proteins are able to bind to their targets – these traits emerge as evolutionary “spandrels", even if they do not confer an intrinsic fitness benefit. Evolutionary trajectories predicted that proteins can evolve strong binding interactions that have no functional role but serve to stabilize the protein if the misfolded state is deleterious. Making connection with the present work, it is possible that multimerization could emerge from these binding interactions, not just to ligand but also to other subunits. If additionally the protein has functional binding, the evolutionary dynamics in [66] predicted that the protein initially gains folding stability but partially loses it as the new binding function develops. Extension of the framework in [66] to include multimerization would similarly probe the interplay between the strength of protein-protein binding and activity.

Understanding the evolution of higher-order protein structure has been an important challenge in evolutionary cell biology. We have presented a computational framework for quantifying the benefit of multimerization in growing and dividing cells in a way that can be directly tied to fitness through the growth rate. The precise fitness benefit is context-dependent, and is determined by the role of the protein of interest in the cellular repertoire, underlying biochemical parameters governing its expression and activity, as well as environmental factors acting as input to its function. As such, multimerization while highly beneficial in one species may be less so in another species. Additionally, mutation bias affecting transcription or translation rates, or substrate binding, may push an enzyme from a regime in which multimers are clearly advantageous to one in which the increase in activity is negligible. This can result in a pressure to decrease the magnitude of ΔG, since the formation of a protein-protein interface involves an increase in surface hydrophobicity and risk of promiscuous interactions. These factors may mitigate any adaptive benefit of multimerization. Ultimately, the interplay between the magnitude of the fitness advantage and the noise inherent in protein numbers will determine the extent to which the forces of selection, mutation, and drift affect the evolutionary trajectory of the emerging PPI in growing and dividing cells.

Appendix A: Substrate Binding and Two-stage Capture

In this Section, we address the mechanism by which enhanced substrate binding may result directly from multimerization, strictly due to interface-interface binding of monomeric subunits and without invoking any further conformational modifications, such as allostery.

In a classic paper, Adam and Delbrück [52] first put forth the hypothesis that the rate at which a surface-bound trap reacts with a substrate diffusing in bulk phase can be enhanced if the substrate first adsorbs to the surface nonspecifically then diffuses in two dimensions before being absorbed by the trap (or equivalently, being modified irreversibly into product). They went on to speculate that this reduction of dimensionality of diffusion, from three to two dimensions, leading to shorter time to capture of reactants by their target sites may have “contributed to the evolutionary advantage of internal membranes.”

Here, we extend this idea to the binding of a diffusing substrate to its target region on the surface of a cytoplasmic protein, present as a monomer or multimeric complex. There is often a significant size difference between reactants and the proteins to which they bind, for example in metabolic pathways, where the substrates are metabolites with a mass of less than 500 Da while the corresponding enzymes are usually about 100 times heavier [67]. Hence, we can think of the reactant as first becoming adsorbed nonspecifically to the surface of the protein through electrostatic, ionic, or hydrophobic interactions, and then diffusing in two dimensions to its binding pocket.

We follow Berg and Purcell’s subsequent re-derivation of Adam and Delbrück’s results and treatment of the effect of target density [53]. The rate of arrival of a substrate molecule to its binding site on the surface of an enzyme is treated as a two-stage process. First, bulk diffusion, characterized by the diffusion constant D in three dimensions, brings substrate molecules to the surface of an enzyme of radius R. Substate molecules are then adsorbed to the surface with mean number $\bar{N}_a$ given by:

$$\bar{N}_a = 4\pi R^2 d\, c_\infty e^{E_a/k_B T}, \tag{A1}$$

where E_a denotes the energy of adsorption, d is given by a molecular interaction distance, and c_∞ is the constant substrate concentration far from the enzyme. The average time to capture, $\bar{t}_c$, of a substrate molecule diffusing on the surface of the enzyme by a binding site of size, s, assumed to be a perfect absorber, is

$$\bar{t}_c = \frac{1.1R^2}{ND'} \ln\left(\frac{1.2R^2}{Ns^2}\right). \tag{A2}$$

where D' is the two-dimensional diffusion constant for the substrate on the surface of the enzyme, and N denotes the number of binding sites (for example, $N = 1, 2$ for monomeric/dimeric forms of the enzyme, respectively).

The Berg-Purcell result is valid under the following assumptions: 1) The equilibration of nonspecifically bound substrate is rapid compared to the rate of absorption by traps[2]; 2) Three dimensional diffusion in the bulk is fast, and therefore in the enzyme-limited regime, the bulk substrate concentration in the vicinity of the enzyme (which in turn determines the surface adsorbed concentration of substrate) is approximately given by c_∞; 3) If the reaction probability is low (or equivalently, many substrate encounters with the target are required for the irreversible reaction from substrate to product to proceed), then the local surface concentration of substrate around each target can approach the constant solute concentration more distant from any target, creating a very shallow depletion zone (the "reaction limit"). In the Michaelis-Menten (MM) kinetic scheme, this condition is equivalent to the assumption that the association/dissociation rates, $k_{\pm 1}$, for substrate-target complex formation are much larger than the catalytic rate, k_{cat}, at which the substrate is irreversibly converted to product.

The average "current" of substrate molecules to their binding sites (or equivalently, the number of substrate molecules absorbed per unit time by perfect absorbing patches on the surface of the enzyme), given by I', is:

$$I' = \bar{N}_a/\bar{t}_c. \tag{A3}$$

If an N-mer is treated as a sphere of radius $R_N = \gamma R_M$, where R_M is the radius of the monomer (for example $\gamma = N^{1/3}$, requiring the volume of a spherical N-mer to be equal to that of N spherical monomers) then the ratio of these currents for the same substrate concentration is

$$\begin{aligned} I'_N/I'_1 &= N\left[\frac{\ln\left(1.2R_M^2/s^2\right)}{\ln\left(1.2R_M^2/s^2\right) + \ln\left(\gamma^2/N\right)}\right] \\ &\approx N\left[1 - \frac{\ln\left(\gamma^2/N\right)}{\ln\left(1.2R_M^2/s^2\right)}\right] \\ &= N\left[1 + \frac{\frac{1}{3}\ln N}{\ln\left(1.2R_M^2/s^2\right)}\right] > N. \end{aligned} \tag{A4}$$

For example, considering dimers and monomers, where the dimer is taken as a sphere with twice the volume of a monomer, then $\gamma = \sqrt[3]{2}$. Taking $s/R_M \sim 0.1$ [70], we have $I'_2/I'_1 = 2.05 > 2$. In the MM scheme, the rate of production formation is given by

$$\frac{dP}{dt} = Nk_{cat}\, c_E^{tot}\, \frac{c_\infty}{c_\infty + K_m} \equiv \nu, \tag{A5}$$

where $K_m = (k_{cat} + k_{-1})/k_1$ is the MM constant, and c_E^{tot} is the total enzyme concentration (in its monomeric or multimeric form). In terms of the current, I'_N, we have: $c_\infty\, k_1 = I'_N/N$. To determine the possible functional advantage of multimerization in terms of the rate of product formation, we can identify two relevant limits: In the first limiting case, if $c_\infty \ll K_m$, and having assumed $k_{\pm 1} \gg k_{cat}$, then $dP/dt \approx k_{cat}\, k_1\, c_E^{tot} c_\infty / k_{-1}$. Therefore, even in the absence of any dimer advantage in stability to degradation (i.e., equal monomer and dimer decay rates, $d_1 = d_2$), where the mean concentration of dimers is expected to be $\bar{c}_D^{tot} = \bar{c}_M^{tot}/2$, we have $\nu_D/\nu_M \approx 2.05/2 > 1$: Dimerization confers a functional advantage to the protein arising purely from diffusion. In the second limiting case, if $c_\infty \gg K_m$, then $dP/dt \approx Nk_{cat}\, c_E^{tot}$, independent of the details of binding kinetics. Hence, absent a stability to degradation (or catalytic) advantage, in this limit, multimers do not present a functional advantage.

Given the crowded nature of the cytopolasm and cost of protein biosynthesis, it is advantageous for enzymes to fulfill their required role in the cellular repertoire – for example, in the case of metabolic enzymes to achieve the requisite metabolic flux – with minimal enzyme concentrations. This is achieved for substrate concentrations high enough to saturate enzymes. Indeed recent work has shown that most measured metabolites in the cell are present at concentrations higher than the K_m for the associated enzyme [54]. However, a downside of maintaining substrate concentrations consistent with the second limiting case considered above is insensitivity to substrate concentration. It has been noted that mechanisms such as allostery can allow for flux regulation in this limit.

Finally, we note that if the current were collected without the aid of surface diffusion, as shown by Berg and Purcell [53]

$$I_N = 4\pi D c_\infty R\, \frac{Ns}{Ns + \pi R} = I_{max}\, \frac{Ns}{Ns + \pi R}, \tag{A6}$$

[2] We note that other works have considered extensions of reduction of dimensionality kinetics for diffusion-limited irreversible targets, where the substrate depletion zones around targets are explicitly treated [68], as well as in the reaction-limited regime for both reversible and irreversible targets [69]. These works show that in both regimes, the rate of two stage capture can depend on the non-target region kinetic rate constants, and not just the equilibrium constant as assumed in the Berg-Purcell result.

where $I_{max} = 4\pi D c_\infty R$ is the diffusive current to a perfectly absorbing sphere of radius R. In the usual limit of small absorbing patches/binding sites, $Ns \ll R$, we have $I_N \approx 4NDsc_\infty$. Hence, without two-stage capture

$$I_N/I_1 = N, \tag{A7}$$

and even in the first limiting case considered, without an advantage in stability to degradation, multimers do not present a functional advantage.

Appendix B: Comparison of Simulations Results With and Without Cell Division With Analytical Distributions of Shahrezaei-Swain [14]

In this Section, we show how the analytical distributions of Shahrezaei-Swain [14], modified according to an effective loss rate due to cell growth and division, agree with simulation results.

Shahrezaei and Swain [14] presented a three-stage model of gene expression in a static cell that captures the stochastic nature of gene activation, transcription, translation, and mRNA and protein decay. The promoter transitions between active and inactive states with rates k_0 and k_1, respectively. Transcription of mRNA occurs with rate v_0 when the promoter is active, and transcripts decay with rate d_0. Translation and decay of protein occur with rates v_1 and d_1, respectively. For proteins that are long-lived compared to mRNA, i.e., $\gamma = d_0/d_1 \gg 1$, the steady-state probability $P(n)$ of there being n proteins in the cell at any time is given by [14]:

$$\begin{aligned} P(n) &= \frac{\Gamma\left(\alpha+n\right)\Gamma\left(\beta+n\right)\Gamma\left(\kappa_0+\kappa_1\right)}{\Gamma\left(n+1\right)\Gamma\left(\alpha\right)\Gamma\left(\beta\right)\Gamma\left(\kappa_0+\kappa_1+n\right)} \\ &\times \left(\frac{b}{1+b}\right)^n \left(1-\frac{b}{1+b}\right)^\alpha \\ &\times {}_2F_1\!\left(\alpha+n, \kappa_0+\kappa_1-\beta, \kappa_0+\kappa_1+n; \frac{b}{1+b}\right), \end{aligned} \tag{B1}$$

where $\alpha = \frac{1}{2}\left(a+\kappa_0+\kappa_1+\phi\right)$, $\beta = \frac{1}{2}\left(a+\kappa_0+\kappa_1-\phi\right)$, and $\phi^2 = \left(a+\kappa_0+\kappa_1\right)^2 - 4a\kappa_0$, with $a = v_0/d_1$, $b = v_1/d_0$, $\kappa_0 = k_0/d_1$, and $\kappa_1 = k_1/d_1$. The mean protein concentration is given by $\bar{n} = abk_0/\left(k_0+k_1\right)$. We modify this result to account for the loss of protein and mRNA due to cell division, where cellular contents are divided approximately equally among daughter cells. This is achieved by replacing the loss rates d_0 and d_1 with $d_0' = d_0 + \lambda$ and $d_1' = d_1 + \lambda$, respectively, where λ is an effective degradation rate due to cell division.

To determine the decay rate due to cell division, λ, we consider the following simplified description of protein expression, as coupled differential equations describing mRNA (m) and protein (n) number dynamics:

$$\begin{aligned} \frac{dm}{dt} &= v_0 - d_0 m - f(t)\frac{m}{2}, && \text{(B2)} \\ \frac{dn}{dt} &= v_1 m - d_1 n - f(t)\frac{n}{2}, && \text{(B3)} \end{aligned}$$

where $f(t)$ is a Dirac comb function representing the loss at division times, at which molecule numbers are (approximately) halved. We define T to be a random variable representing the cell cycle length, and assume that $n(T_-) = 2n(0)$ and $n(t = T_+) = n(0)$ at steady state, where $t = T_-$ represents the instant in time just before, and $t = T_+$ is the instant in time just after the cell division. Averaging over the cell cycle, where for a periodic signal $\langle dn/dt \rangle = 0$, Eq. B3 becomes

$$\begin{aligned} v_1 \langle m \rangle - d_1 \langle n \rangle - \left\langle f(t)\frac{n}{2} \right\rangle \\ &= v_1 \langle m \rangle - d_1 \langle n \rangle - \frac{n(T)}{2T}, \\ &= v_1 \langle m \rangle - \left(d_1 + \lambda\right) \langle n \rangle, \end{aligned} \tag{B4}$$

where

$$\lambda(T) = \frac{n(T)}{2T\langle n \rangle} = \frac{d_0 d_1 \left[d_0 Y \left(1 - e^{d_1 T}\right) - d_1 \left(1 - e^{d_0 T}\right) Z\right]}{d_1^2 \left(1 - e^{d_0 T}\right) Z - d_0 d_1^2 T Y Z + d_0^2 Y \left(d_1 T Z - \left(1 - e^{d_1 T}\right)\right)}, \tag{B5}$$

with similar results from Eq. B2. In the above, $Y = 1 - 2e^{d_0 T}$, $Z = 1 - 2e^{d_1 T}$. The solutions $m(t)$ and $n(t)$ are shown in Fig. S1, plotted against the mRNA and protein numbers from simulation data averaged over cells. As shown in Fig. S2, we note that $\lambda(T)$ given by Eq. B5 agrees closely with the putative expression $\lambda_0(T) = \ln 2/T$, which represents the loss due to cell division with half-life equal to the doubling time of the cells [22, 71, 72]. Importantly, we note from Eq. B4 that effective decay rates can be defined as $d' = d + \lambda$ to capture loss due to cell division.

Using effective decay rates, the analytical probability distribution of protein number from Shahrezaei and Swain [14] absent cell division (Eq. B1) can be modified according to a given value of λ, giving $P(n|\lambda)$. The effective loss rate due to division, $\lambda(T)$, depends on the cell cycle duration, T; the distribution of λ, shown in Fig. S3, is determined numerically from the simulated distribution of T (Fig. S4B), obtained from the model of cell growth and division with cell size control used in this work [17].

From these, we can construct

$$P(n) = \sum_{\lambda} P(n|\lambda)P(\lambda), \tag{B6}$$

Additionally, stochastic partitioning of cellular contents is modeled by introducing the binomial random variable X, whose value x represents the fraction of proteins inherited from the previous cell division

$$B(x|n) = \binom{2n}{xn} 2^{-2n}. \tag{B7}$$

The final protein number distribution accounting for noisy partitioning is then given by that of the random variable $N' = N \times X$, , shown in Fig. S6 (green line). We note good agreement between the modified analytical result of Shahrezaei and Swain [14] and the distribution of protein number from simulations with cell division.

A rigorous method of incorporating binomial partitioning would include in the master equation approach of [14] an explicit loss term, as in Eq. B4 modified with a binomial random variable with mean equal to 1/2. While this formulation is beyond the scope of the present analysis, recent work has addressed molecule number distributions in a population of exponentially growing and dividing cells taking into account its age structure [73].

Acknowledgments

Support was provided by the Multidisciplinary University Research Initiative awards W911NF-09-1-0444 and W911NF-14-1-0411 from the US Army Research Office (ML), National Institutes of Health award R01-GM03827 (ML), National Science Foundation award MCB-1050161 (ML), and National Science Foundation Award PHY-0645652 (SS).

[1] K. Hashimoto, H. Nishi, S. Bryant, and A. R. Panchenko, Physical Biology **8**, 035007 (2011).

[2] M. H. Ali and B. Imperiali, Bioorganic & Medicinal Chemistry **13**, 5013 (2005).

[3] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, Journal of Molecular Biology **112**, 535 (1977).

[4] J.-P. Changeux, Annual Review of Biophysics **41**, 103 (2012).

[5] M. Lynch, Proceedings of the National Academy of Sciences **110**, E2821 (2013).

[6] M. B. Elowitz, Science **297**, 1183 (2002).

[7] P. Thomas, N. Popović, and R. Grima, Proceedings of the National Academy of Sciences **111**, 6994 (2014).

[8] S. Bershtein, A. W. R. Serohijos, S. Bhattacharyya, M. Manhart, J.-M. Choi, W. Mu, J. Zhou, and E. I. Shakhnovich, PLOS Genetics **11**, e1005612 (2015).

[9] J. V. Rodrigues, S. Bershtein, A. Li, E. R. Lozovsky, D. L. Hartl, and E. I. Shakhnovich, Proceedings of the National Academy of Sciences **113**, E1470 (2016).

[10] B. V. Adkar, M. Manhart, S. Bhattacharyya, J. Tian, M. Musharbash, and E. I. Shakhnovich, bioRxiv p. 088013 (2017).

[11] A. J. Waite, N. W. Frankel, Y. S. Dufour, J. F. Johnston, J. Long, and T. Emonet, Molecular Systems Biology **12**, 895 (2016).

[12] A. Amir and J. Lin, Cell Systems **5**, 358 (2017).

[13] M. Hashimoto, T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto, Proceedings of the National Academy of Sciences **113**, 3251 (2016).

[14] V. Shahrezaei and P. S. Swain, Proceedings of the National Academy of Sciences **105**, 17256 (2008).

[15] S. H. Northrup and H. P. Erickson, Proceedings of the National Academy of Sciences **89**, 3338 (1992).

[16] N. E. Buchler, U. Gerland, and T. Hwa, Proceedings of the National Academy of Sciences **102**, 9559 (2005).

[17] M. Osella, E. Nugent, and M. C. Lagomarsino, Proceedings of the National Academy of Sciences **111**, 3431 (2014).

[18] A. Amir, Physical Review Letters **112**, 208102 (2014).

[19] S. Taheri-Araghi, S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun, Current Biology **25**, 385 (2015).

[20] J. M. Guberman, A. Fay, J. Dworkin, N. S. Wingreen, and Z. Gitai, PLoS computational biology **4**, e1000233 (2008).

[21] D. T. Gillespie, The Journal of Physical Chemistry B **113**, 1640 (2009).

[22] R. Marathe, V. Bierbaum, D. Gomez, and S. Klumpp, Journal of Statistical Physics **148**, 608 (2012).

[23] K. Nath and A. L. Koch, Journal of Biological Chemistry **245**, 2889 (1970).

[24] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, Nature Genetics **31**, 69 (2002).

[25] H. Salman, N. Brenner, C.-k. Tung, N. Elyahu, E. Stolovicki, L. Moore, A. Libchaber, and E. Braun, Physical Review Letters **108**, 238105 (2012).

[26] N. Brenner, E. Braun, A. Yoney, L. Susman, J. Rotella, and H. Salman, The European Physical Journal E **38**, 102 (2015).

[27] G. Baumann, M. W. Stolar, and T. A. Buchanan, Endocrinology **119**, 1497 (1986).

[28] D. S. Goodsell and A. J. Olson, Annual Review of Biophysics and Biomolecular Structure **29**, 105 (2000).

[29] T. Perica, J. A. Marsh, F. L. Sousa, E. Natan, L. J. Colwell, S. E. Ahnert, and S. A. Teichmann, Biochemical Society Transactions **40**, 475 (2012).

[30] J. A. Marsh, H. Hernández, Z. Hall, S. E. Ahnert, T. Perica, C. V. Robinson, and S. A. Teichmann, Cell **153**, 461 (2013).

[31] Y. Taniguchi, P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, Science **329**, 533 (2010).

[32] E. Krissinel and K. Henrick, Journal of Molecular Biology **372**, 774 (2007).

[33] J. Zhang, S. Maslov, and E. I. Shakhnovich, Molecular Systems Biology **4** (2008).

[34] C. F. Reboul, B. T. Porebski, M. D. W. Griffin, R. C. J. Dobson, M. A. Perugini, J. A. Gerrard, and A. M. Buckle, PLoS computational biology **8**, e1002537 (2012).

[35] E. D. Levy, E. B. Erba, C. V. Robinson, and S. A. Teichmann, Nature **453**, 1262 (2008).

[36] N. Brooijmans, K. A. Sharp, and I. D. Kuntz, Proteins: Structure, Function, and Genetics **48**, 645 (2002).

[37] O. Jardine, J. Gough, C. Chothia, and S. A. Teichmann, Genome research **12**, 916 (2002).

[38] M. D. W. Griffin and J. A. Gerrard, in *Protein Dimerization and Oligomerization in Biology* (Springer New York, New York, NY, 2012), pp. 74–90.

[39] K. Helmstaedt, S. Krappmann, and G. H. Braus, Microbiology and Molecular Biology Reviews **65**, 404 (2001).

[40] J. Kuriyan and D. Eisenberg, Nature **450**, 983 (2007).

[41] K. A. Reynolds, R. N. McLaughlin, and R. Ranganathan, Cell **147**, 1564 (2011).

[42] J. B. Christensen, T. P. S. da Costa, P. Faou, F. G. Pearce, S. Panjikar, and M. A. Perugini, Scientific Reports **6**, 37111 (2016).

[43] M. D. W. Griffin, R. C. J. Dobson, F. G. Pearce, L. Antonio, A. E. Whitten, C. K. Liew, J. P. Mackay, J. Trewhella, G. B. Jameson, M. A. Perugini, et al., Journal of Molecular Biology **380**, 691 (2008).

[44] M. Kafri, E. Metzl-Raz, G. Jona, and N. Barkai, Cell Reports **14**, 22 (2016).

[45] M. Lynch and G. K. Marinov, Proceedings of the National Academy of Sciences **112**, 15690 (2015).

[46] E. Dekel and U. Alon, Nature **436**, 588 (2005).

[47] D. A. Charlebois, Physical Review E **92**, 022713 (2015).

[48] G.-W. Li, D. Burkhardt, C. Gross, and J. S. Weissman, Cell **157**, 624 (2014).

[49] E. Noor, A. Flamholz, A. Bar-Even, D. Davidi, R. Milo, and W. Liebermeister, PLoS computational biology **12**, e1005167 (2016).

[50] D. J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans, Nature **514**, 376 (2014).

[51] A. Weiss and J. Schlessinger, Cell **94**, 277 (1998).

[52] G. Adam and M. Delbrück, *Reduction of dimensionality in biological diffusion processes* (Structural chemistry and molecular biology, 1968).

[53] H. C. Berg and E. M. Purcell, Biophysical Journal **20**, 193 (1977).

[54] B. D. Bennett, E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien, and J. D. Rabinowitz, Nature Chemical Biology **5**, 593 (2009).

[55] S. Bershtein, W. Mu, and E. I. Shakhnovich, Proceedings of the National Academy of Sciences **109**, 4857 (2012).

[56] M. v. Smoluchowski, Kolloid-Zeitschrift **21**, 98 (1917).

[57] N. J. Marianayagam, M. Sunde, and J. M. Matthews, Trends in Biochemical Sciences **29**, 618 (2004).

[58] M. Lynch, Molecular Biology and Evolution **29**, 1353 (2012).

[59] T. Vavouri, J. I. Semple, R. Garcia-Verdugo, and B. Lehner, Cell **138**, 198 (2009).

[60] A. Y. Weiße, D. A. Oyarzún, V. Danos, and P. S. Swain, Proceedings of the National Academy of Sciences **112**, E1038 (2015).

[61] P. T. Beernink and D. R. Tolan, Proceedings of the National Academy of Sciences **93**, 5374 (1996).

[62] T. Perica, Y. Kondo, S. P. Tiwari, S. H. McLaughlin, K. R. Kemplen, X. Zhang, A. Steward, N. Reuter, J. Clarke, and S. A. Teichmann, Science **346**, 1254346 (2014).

[63] B. K. Shoichet, W. A. Baase, R. Kuroki, and B. W. Matthews, Proceedings of the National Academy of Sciences **92**, 452 (1995).

[64] X. Wang, G. Minasov, and B. K. Shoichet, Journal of molecular biology **320**, 85 (2002).

[65] M. R. Mitchell, T. Tlusty, and S. Leibler, Proceedings of the National Academy of Sciences **113**, E5847 (2016).

[66] M. Manhart and A. V. Morozov, Proceedings of the National Academy of Sciences **112**, 1797 (2015).

[67] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, Nucleic Acids Research **38**, D750 (2009).

[68] D. Wang, S.-Y. Gou, and D. Axelrod, Biophysical chemistry **43**, 117 (1992).

[69] D. Axelrod and M. D. Wang, Biophysical Journal **66**, 588 (1994).

[70] J. Liang, C. Woodward, and H. Edelsbrunner, Protein Science **7**, 1884 (1998).

[71] V. Bierbaum and S. Klumpp, Physical Biology **12**, 066003 (2015).

[72] M. Osella and M. C. Lagomarsino, Physical Review E **87**, 012726 (2013).

[73] P. Thomas, Journal of The Royal Society Interface **14**, 20170467 (2017).

[74] See Supplemental Material at [URL will be inserted by publisher] for supporting figures.

TABLE II: Model predictions based on experimental data. Solutions to Eqs. 7-9 are compared with experimental data for a selection of *E. coli* genes. Mean total protein ($\overline{P}$) and mRNA ($\overline{m}$) counts are taken from [31]. ($\overline{m}$ is determined by the value of the composite parameter $P_{on}v_0/d_0d_1$, so the values of the individual variables were not considered.) Interface ΔG values are taken from PDBePISA and are given in kcal/mol, with ΔG_1 representing the first (dimer-forming) interface and ΔG_2, if applicable, being the secondary (tetramer-forming) interface. Translation rate v_1 was chosen iteratively in order to yield a calculated protein level sufficiently close to the literature value given in [31]. For this value of v_1, the percentage of monomers (M), dimers (D), and tetramers (T) were obtained from Eqs. 7-9 and compared to the assembly recorded in the PDB/PDBePISA. *metK* and *gatZ* are listed as dimeric in the PDB but analyzed as likely tetramers by PDBePISA.

Gene Name	PDB ID	$\overline{P}$	$\overline{m}$	ΔG_1	ΔG_2	Assembly	M	D	T
map	1c22	125	1.64	-3.4	-1.6	Monomer	**100.0**	0.0	0.0
fabD	2g2o	5	0.36	-2.4	-2.4	Monomer	**100.0**	0.0	0.0
pgk	1zmr	564	0.70	-1.1	-	Monomer	**100**	0.0	0.0
cspA	1mjc	715	0.39	0.3	-	Monomer	**100.0**	0.0	0.0
livJ	1z16	8	0.44	-1.1	-0.2	Monomer	**100.0**	0.0	0.0
acnB	1l5j	232	0.06	-4.7	-	Monomer	**100.0**	0.0	0.0
yhbY	1ln4	137	0.28	-4.7	-	Monomer	**100.0**	0.0	0.0
adk	1ank	687	0.09	-3.7	-	Monomer	**100.0**	0.0	0.0
pstS	2abh	3	2.2	-1.6	-	Monomer	**100.0**	0.0	0.0
ybbN	3qou	128	0.09	-6.1	0	Monomer	**99.8**	0.2	0.0
hdeA	1bg8	21	6.08	-22.9	-6.7	Dimer	0.4	**99.2**	0.4
csrA	1y00	474	0.94	-24.9	-	Dimer	0.2	**99.8**	0.0
hdeB	2xuv	8	1.5	-21.2	-7.7	Dimer	11.5	**88.5**	0.0
thrS	4p3p	784	0.82	-17.7	-1.4	Dimer	0.1	**99.9**	0.0
ppiB	2nul	385	0.29	-17.7	-	Dimer	0.3	**99.7**	0.0
fabA	1mka	611	0.51	-16.8	2.9	Dimer	0.2	**99.8**	0.0
tktA	2r8o	290	0.10	-32	-	Dimer	0.3	**99.7**	0.0
aceE	2qtc	843	0.78	-32.6	-1.8	Dimer	0.1	**99.9**	0.0
ybeX	4hg0	44	0.07	-22.6	-5.2	Dimer	2.2	**97.8**	0.0
serC	1bjn	188	0.39	-23.0	0.1	Dimer	0.5	**99.5**	0.0
purA	2gcq	310	1.09	-12.8	0.6	Dimer	5.3	**94.6**	0.0
pflB	1mzo	239	0.39	-12.3	1.7	Dimer	9.1	**90.9**	0.0
tpiA	1tmh	768	0.27	-14.4	-3.8	Dimer	0.9	**99.1**	0.0
aspC	1bqa	248	0.31	-21.5	-	Dimer	0.4	**99.6**	0.0
fba	1b57	676	0.21	–30.4	-	Dimer	0.1	**99.9**	0.0
gpmA	1e58	760	1.08	-4.9	-	Dimer*	**99.9**	0.1	0.0
glf	1i8t	193	0.07	-3.7	-0.6	Dimer*	**100.0**	0.0	0.0
metK	1xra	347	0.58	-19.5	-7.0	Dimer/Tetramer	0.3	**98.7**	1.1
gatZ	2fiq	290	0.40	-10.0	-12.0	Dimer/Tetramer	22.5	12.8	**64.8**
fabI	1c14	342	0.29	-14.9	-11.9	Tetramer	0.5	14.6	**84.9**
tnaA	2c44	351	0.09	-16.4	-15	Tetramer	0.3	1.4	**98.3**
gapA	1gae	2380	2.11	-28.3	-12.8	Tetramer	0	2.7	**97.2**
secB	1qyn	673	0.38	-6.1	-10.2	Tetramer*	**99.0**	1.0	0.0