Lessons from equilibrium statistical physics regarding the assembly of protein complexes

Pablo Sartori[a,b,c,1] and Stanislas Leibler[a,b,c,1]

[a]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540; [b]Laboratory of Living Matter, The Rockefeller University, New York, NY 10065; and [c]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10065

Contributed by Stanislas Leibler, November 22, 2019 (sent for review June 27, 2019; reviewed by Christopher Jarzynski and Johan Paulsson)

Cellular functions are established through biological evolution, but are constrained by the laws of physics. For instance, the physics of protein folding limits the lengths of cellular polypeptide chains. Consequently, many cellular functions are carried out not by long, isolated proteins, but rather by multiprotein complexes. Protein complexes themselves do not escape physical constraints, one of the most important being the difficulty of assembling reliably in the presence of cellular noise. In order to lay the foundation for a theory of reliable protein complex assembly, we study here an equilibrium thermodynamic model of self-assembly that exhibits 4 distinct assembly behaviors: diluted protein solution, liquid mixture, "chimeric assembly," and "multifarious assembly." In the latter regime, different protein complexes can coexist without forming erroneous chimeric structures. We show that 2 conditions have to be fulfilled to attain this regime: 1) The composition of the complexes needs to be sufficiently heterogeneous, and 2) the use of the set of components by the complexes has to be sparse. Our analysis of publicly available databases of protein complexes indicates that cellular protein systems might have indeed evolved so as to satisfy both of these conditions.

self-assembly | protein complex

Protein complexes are assembled with high compositional accuracy, evidenced, for example, by the possibility of crystallization of complexes as large as the ribosome (1). This is remarkable, because, during assembly, a growing complex has to discriminate its specific components from a multicomponent mixture of hundreds of different protein species that are part of the proteome. Failure to solve this discriminatory task could result in assembly of chimeric structures composed of fragments from different complexes, impairing normal cellular function (2).

Assembly of protein complexes can also be viewed as a second stage of creating functional cellular structures, the first being the assemblage of amino acids into proteins, achieved by ribosomes. A modest alphabet of 20 amino acids encodes thousands of different proteins. Proteins typically contain all 20 amino acids, so that the amino acid usage by proteins is "dense" rather than "sparse." Nature, furthermore, reuses amino acids many times within the same protein, which makes the compositional heterogeneity of each protein low. This can be contrasted with the assembly of complexes, which seem to use proteins sparsely, so that each complex contains only a small fraction of the available proteome. At the same time, complexes are often highly heterogeneous, that is, composed of many different protein species (3).

The sparsity and heterogeneity of complexes should come as a surprise, as they imply that the proteome might not be exploited in combinatorial manner. Indeed, the vast repertoire of hundreds of proteins is combined to result in a comparable number of complexes (Fig. 1). This suggests that "combinatorial expansion" of proteins into complexes does not occur generically, and may instead be restricted to particular functions, such as regulation or signaling (4, 5). In these cases, proteins participate in several complexes; for example, cyclin-dependent kinases can be part of several cell cycle regulatory complexes (6). Proteins can have specific interactions with many partners, a phenomenon known as *promiscuity*. The promiscuity of proteins may potentially result in the formation of disordered chimeric structures. For example, a single point mutation is sufficient to create a novel protein–protein interaction, which can result in chimeric assembly of proteins (7). Notwithstanding these challenges, protein complexes typically assemble from their constituents accurately and carry out cellular functions with remarkable speed and precision (8).

Elucidating the characteristics of protein complexes that enable them to assemble reliably, and studying how these characteristics affect the organization of the proteome, can be viewed as fundamental goals of cell biology. Recently, there have been significant advances toward achieving these goals, due to the progress in experiments (7, 9, 10), bioinformatics (11, 12), and molecular dynamics simulations (13). However, a general theoretical framework to understand protein complex formation and usage is still lacking. One major difficulty in developing such a framework is the large diversity of cellular protein complexes. Some complexes, such as microtubules, exhibit unbounded growth (14). Others, such as ribosomes, have a well-defined finite size (1). To complicate matters further, the latter complexes can be further divided among those that exhibit strong symmetries, such as the bacterial flagellar motor (15), and those that are fully asymmetric, such as ribosomes (1). Whereas the principles of assembly of many symmetric complexes have been studied (12), the same is not true for asymmetric complexes.

Significance

In order to carry out their functions, most proteins assemble into multicomponent complexes. In the process of assembly, complexes need to discriminate their specific components from a mixture of hundreds of different proteins present in the cell. To assess some of the implications of this requirement, we develop a minimal model of self-assembly based on equilibrium statistical physics. We argue that the need to assemble reliably imposes fundamental constraints on the characteristics of complexes, which we support with analysis of available structural and compositional data. Our work constitutes only a step toward future theory of protein complex assembly, which will have to incorporate also nonequilibrium and kinetic aspects of this fundamental and rich, yet theoretically neglected, problem.

Author contributions: P.S. and S.L. designed research; P.S. performed research; and P.S. and S.L. wrote the paper.

Reviewers: C.J., University of Maryland; and J.P., Harvard University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: psartori@rockefeller.edu or livingmatter@rockefeller.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1911028117/-/DCSupplemental.

First published December 23, 2019.

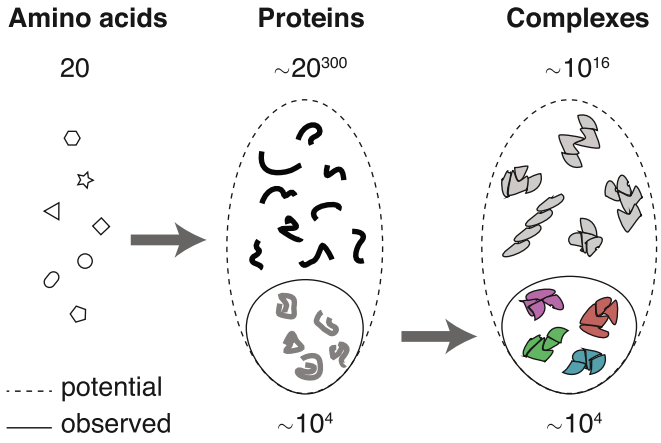

Fig. 1. Usage of amino acids by proteins, and of proteins by complexes. The typical length of proteins (~ 300 residues) is largely limited by folding (38). Twenty amino acids (geometric shapes) potentially thus encode $\sim 20^{300}$ proteins (black lines). Although the observed repertoire of proteins (gray lines) is much smaller, for example, $\sim 10^4$ for *Saccharomyces cerevisiae* (39), there is a clear combinatorial expansion from amino acids to proteins. Contrary to this, the observed number of complexes, with the reported average of 4 different proteins per complex (40) (colored shapes), is comparable to the number of proteins, without a trace of combinatorial expansion.

The aim of this article is to begin to develop a theoretical framework, which could ultimately be applied to (asymmetric) protein complexes, by extending a recent model of self-assembly (16). As will be elaborated in *Discussion*, cellular assembly of protein complexes can be a highly controlled, nonequilibrium kinetic process. Still, we will constrain our present theoretical study to equilibrium statistical physics alone and explore what constraints thermodynamics imposes on assembly of complexes. Interestingly, we shall see that these constraints alone can—at least partly—explain the observed heterogeneity of asymmetric complexes and their sparse usage of the proteome. We will also analyze existing structural, compositional, and interaction data of protein complexes to further evaluate some biological implications of our theoretical findings.

Results

Multifarious Mixtures of Components Exhibit 4 Assembly Regimes. In our model, protein-like components form a multicomponent mixture. When 2 components are in close proximity, they can interact specifically. We specify the interactions of components via the complexes of which the components are part. In particular, if 2 components are bound to each other as part of the same complex, we assume they can interact specifically with binding energy E. Conversely, we assume that components not forming part of the same complex have a null binding energy. Such components still can interact nonspecifically, provided their concentration, p, is large. This model has been formulated and studied previously in ref. 16. We extend its analysis to allow for variable heterogeneity and sparsity. A detailed account of the model is presented in *Materials and Methods*.

Just like changing the temperature and pressure of a gas can turn it into a liquid, changing the binding energy, E, and the chemical potential, μ, of the component mixture can fundamentally alter its properties. For an ideal dilute mixture, the chemical potential μ is given by $\mu = k_{\mathrm{B}}\,T\log(p)$, where p is the concentration of the components relative to the solvent, k_{B} is Boltzmann's constant, and T is the temperature (hereafter, we express energy in units of $k_{\mathrm{B}}\,T$). As shown in Fig. 2 (which describes a lattice implementation of our model; see also *SI Appendix*, section A), for low negative values of the chemical potential, the mixture is in a dilute solution (DS) regime, in which components interact only transiently with each other. This is the regime in which biochemical reactions have been traditionally studied (17). If, on the other hand, the chemical potential is high but the binding energy is low, the mixture increases its density and behaves as a liquid (L). The properties of this liquid are somewhat similar to those of droplets observed at cellular scale (18, 19). These droplets are membraneless organelles composed, among other components, of many proteins, and are believed to form by simple liquid–liquid phase separation. Finally, if both the chemical potential and the binding energy are high, the liquid mixture changes into a "chimeric" (Ch) regime, in which fragments of several protein complexes bind in an unruly manner to each other. This regime can evoke cellular inclusion bodies, where overexpressed recombinant proteins form disordered solid aggregates (20). These 3 regimes are conceptually close to phases of inert materials, and will not be discussed here further (see, however, *SI Appendix*, sections A to D).

We shall rather focus our attention on a regime more relevant for understanding protein complexes, which arises when the values of binding energy and chemical potential are comparable. In this "multifarious assembly" (MA) regime, a large protein complex can self-assemble accurately, for example, starting (nucleating) from a small fragment of such complex (nucleation seed) (Fig. 2*A*) (16). As shown in Fig. 2 *E* and *F*, the MA regime is bounded by the L and Ch regimes. The range of concentrations $p_{\max}/p_{\min}$ in which reliable assembly is possible is given by (*SI Appendix*, section B)

$$p_{\max}/p_{\min} \approx \exp\left(E + \mu_0\right), \qquad [1]$$

where E is the binding energy used to discriminate between specific and nonspecific interactions, and $\mu_0 < 0$ is a reference chemical potential that depends on the characteristics of the mixture. Eq. **1** implies that a binding energy of a few $k_{\mathrm{B}}\,T$ is sufficient to ensure reliable assembly in a range of concentrations spanning several orders of magnitude. The parameter μ_0 depends logarithmically on the number of different complexes, K, and the number of component types, or "species," N_{tot}. Furthermore, it also depends on the characteristics of a typical complex c and a typical component species α. For a complex c, these characteristics are the total number of components that the complex contains (i.e., its size), M_c, and the number of different species among the components of the complex, $N_c \leq M_c$. We simply characterize a component of species α by the number of "binding links" that it can establish, z_α. For the sake of simplicity, we will limit ourselves in the following to the case $z_\alpha = 4$, in which each component can establish 4 binding links (this is well suited for numerical simulations on a squared lattice). However, our results can be generalized to other values of z_α; see *SI Appendix*, section C.

Heterogeneity and Sparsity Constrain Reliable Assembly. The accurate assembly of complexes is a daunting discriminatory task. Proteins accomplish this task because their interactions and the composition of complexes they form are the result of "constrained evolution." That is, the characteristics of complexes have evolved to ensure their cellular function, while, at the same time, they have been constrained to assemble reliably. As argued above, an important quantity, which is closely related to the reliability of the assembly process, is protein promiscuity: If the promiscuity of a protein were exceedingly high, undesired protein species could interfere with this protein during the assembly process, which would then result in the formation of "chimeric structures." Therefore, we expect that evolution tuned protein promiscuity so that proteins do not form such nonfunctional chimeras.

In our model, the promiscuity of a protein-like component is related to the number of different complexes, K, and to their

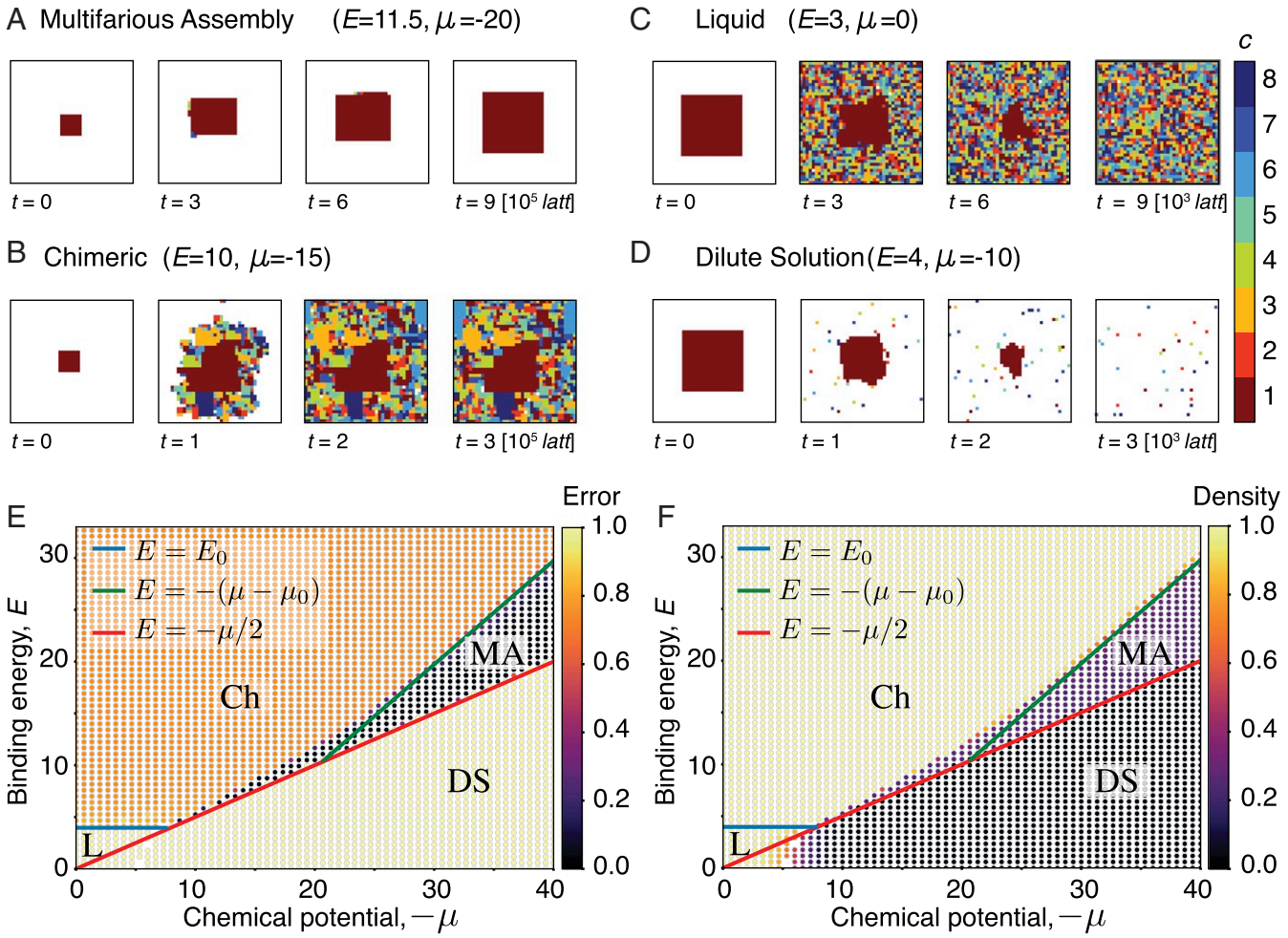

Fig. 2. Four regimes of a multicomponent mixture. In our simplified equilibrium statistical mechanics model, individual different components, $N_{\mathrm{tot}} = 400$ (small, unit cell-sized, squares), have specific interactions with their neighbors and can form $K = 8$ different complexes (large squares) of size $M_c = 400$, each represented by a different color (see color bar). Colors of complexes are assigned to individual components based on whether their neighbors in the lattice are also neighbors in the given complex. Random colors are assigned to resolve ambiguities, and white lattice sites denote absence of any component (see *SI Appendix*, section A for a detailed description of the coloring algorithm). (*A*) In the MA regime, a small fragment of a complex (small brown square at $t = 0$) can be used to nucleate the assembly of the whole complex ($c = 1$, large brown square). (*B*) In the Ch regime, the same small fragment nucleates a disordered aggregate from fragments of many complexes. (*C*) In the L regime, a whole well-assembled complex will be unstable and will "melt" into a dense fluid-like mixture, in which small fragments of complexes are constantly being rearranged (unlike in Ch, where there are no rapid rearrangement dynamics). (*D*) In the DS regime, the initial full complex dissolves quickly into a set of separate proteins. Only very small transient clusters of proteins can form. (*E* and *F*) The 4 regimes observed in *A*–*D* correspond to distinct values of the chemical potential and binding energy, and can be determined by the assembly error (portion of the correctly assembled complex) and the density of components. In *SI Appendix*, section B, we characterize the boundaries that separate the regimes. Note that each circle in these graphs corresponds to separate Monte Carlo simulations, in which the error of assembly is evaluated. See *Materials and Methods* for detailed description of the parameters used in this and other figures.

characteristics, such as their compositional heterogeneity, $h_c \equiv N_c/M_c$, and their sparsity, $a_c \equiv (N_{\mathrm{tot}} - N_c)/N_{\mathrm{tot}}$, that is, the fraction of their proteome usage. One can show that the promiscuity of a component species α scales as (see *SI Appendix*, section C for a detailed derivation)

$$\pi_\alpha \approx K(1 - a_c)/h_c. \qquad [2]$$

One can then express the evolutionary constraint of reliable self-assembly by relating the probability of the formation of chimeric structures to component promiscuity, and requiring that this probability is negligible. By doing so, we find that the number of possible coexisting complexes, their heterogeneity, their size, and their sparsity obey an important constraint relation,

$$K \lesssim h_c^{3/2} M_c^{1/2} (1 - a_c)^{-3/2}, \qquad [3]$$

where the values of the exponents are given for $z_\alpha = 4$; see *SI Appendix*, section C for generalization. Eq. **3** provides the scaling for the surface of transition between the MA regime, in which chimeric structures are avoided, and the Ch regime, in which they readily form (depicted in Fig. 3*A*). Although this relation holds in the limit $M_c \to \infty$, our analysis of Monte Carlo simulations is compatible with Eq. **3** (however, to unequivocally determine the scaling would require a much larger range of complex sizes; *SI Appendix*, section D).

The key determinants of the transition from the MA regime to the Ch regime, predicted by the constraint of Eq. **3**, have direct implications for the composition of complexes and organization of their components. Eq. **3** can be understood as a trade-off between the number of coexisting complexes, their heterogeneity, and their sparsity: More complexes can coexist if they are more heterogeneous, and if they make a sparser usage of the proteome. To see that this is indeed the case, let us first

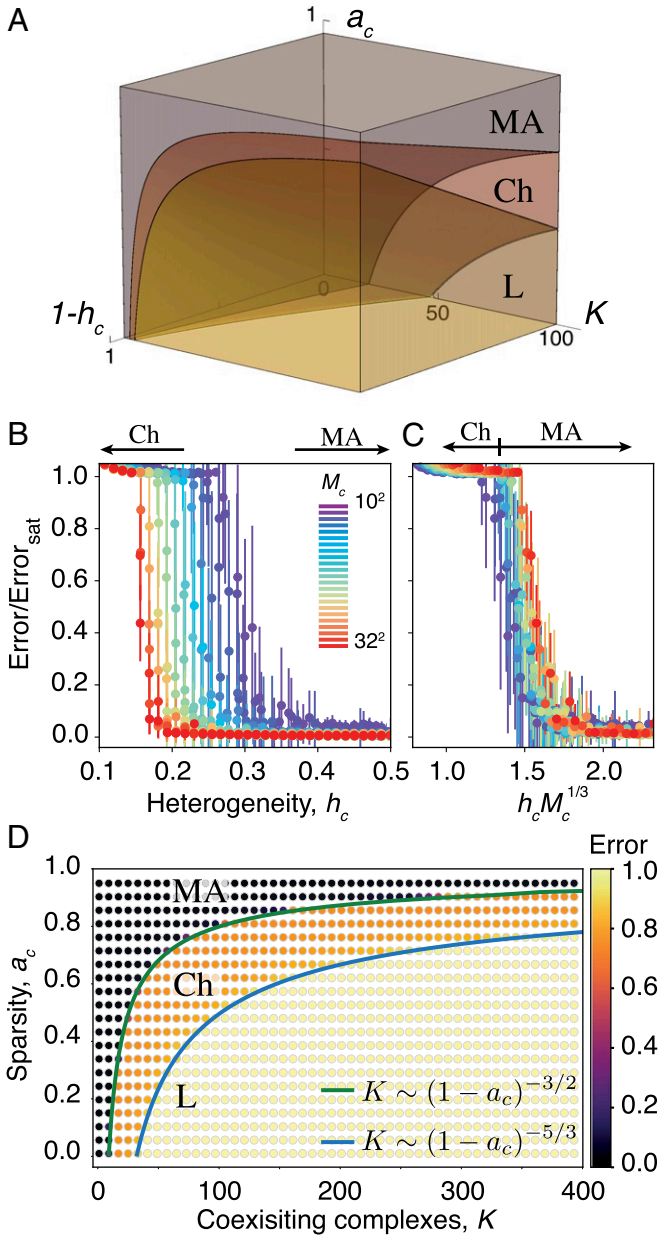

Fig. 3. Characteristics of complexes shape the MA regime. (*A*) Schematic representation of how the different regimes depend on the number of complexes K, their heterogeneity h_c, and their sparsity a_c. Sparsity greatly increases the number of complexes that can be reliably assembled by expanding MA. (*B*) As the heterogeneity of a complex increases, its assembly error decreases (normalized by a saturation error). The system thus transitions from Ch to MA. (*C*) Data collapse of the data in *B* using $h_c M_c^{1/3}$; see also *SI Appendix*, section D. (*D*) Cross-section of the phase diagram sketched in *A*. As the sparsity increases, the number of coexisting complexes diverges algebraically.

consider a single complex type ($K = 1$) prepared in a mixture that contains only its constituent components, so that $N_{\rm tot} = N_c$ and $a_c = 0$. Eq. **3** constrains the heterogeneity of the complex to be larger than a quantity that scales as a power of its size, $h_c \gtrsim M_c^{-1/3}$. Therefore, by increasing the heterogeneity of the complexes, one crosses a transition from the Ch regime to the desired MA regime, as shown in Fig. 3 *B* and *C*. This mechanism might thus explain the observed high heterogeneity among cellular protein complexes as the means of avoiding assembly of incorrect chimeric structures.

Next, let us consider the possibility of combinatorial usage of components in different complexes. In the case of a dense usage of the set of components, that is, $a_c = 0$, the reliable assembly constraint, Eq. **3**, implies that the number of possible coexisting complexes is $K \lesssim h_c N_{\rm tot}^{1/2}$. Such increase of the number of complexes with increasing number $N_{\rm tot}$ of component species implies that combinatorial usage of components in complexes is indeed possible (16). However, the reliability constraint makes the combinatorial aspect only sublinear, and therefore weak: An increase by a factor of 100 in the number of component species merely increases by 10 the number of possible complexes. Therefore, from a biological perspective, reliable assembly introduces a constraint that vastly reduces the possibilities of combinatorial expansion from proteins into protein complexes.

In order for many complexes to coexist, an alternative to this weak combinatorial usage is needed. This alternative is achieved by letting complexes make a sparse usage of the set of components, that is, $a_c \lesssim 1$. To see this, note that the number of possible coexisting complexes, K, diverges in Eq. **3** as the component usage becomes more and more sparse, $a_c \to 1$; see also Fig. 3*D*. An important consequence of such behavior is that the number of coexisting complexes scales superlinearly with the number of component species when a sparse usage is allowed,

$$K \approx N_{\rm tot}^{3/2} M_c^{-1}. \qquad [4]$$

From a biological perspective, due to the previously evoked sparsity of the proteome usage, an increase by a factor of 100 in the number of component species results thus in an increase by a factor of a 1,000 in the number of complexes. Therefore, a sparse usage of the proteome may indeed help to insure the observed coexistence of many different types of protein complexes within the cell.

Overall, we have shown that, in order to avoid chimeric assembly, the composition of complexes and organization of their components must be such that Eq. **3** is satisfied. This implies that many complexes can coexist only if they have a heterogeneous composition and make a sparse usage of their components. In the following, we discuss the biological implications of these findings.

Structural and Compositional Data Point toward High Heterogeneity of Protein Complexes. How relevant are the theoretical arguments presented above for cellular protein complexes? It is clear that the proteome usage is generically sparse: Even a complex with as many different protein species as the ribosome, for which $N_c \lesssim 10^2$, contains only a small fraction of the proteome, $N_{\rm tot} \gtrsim 10^3$. This gives a lower bound for the sparsity of complexes: $a_c > 0.9$.

To address the issue of heterogeneity, in particular, whether highly heterogeneous complexes might coexist more easily, we analyzed a publicly available database of protein complexes (21). The resulting histogram of the heterogeneity of these complexes, separated by their size (see *SI Appendix*, section H for details), is depicted in Fig. 4*A*. The histogram reveals a large abundance of high-heterogeneity complexes, $h_c \in [0.8, 1.0]$, which supports the arguments presented in this work. From Fig. 4*A*, it is also apparent that a significant number of complexes have intermediate values of heterogeneity, $h_c \in [0.4, 0.6)$.

Given that our theory applies to structures without any geometrical symmetry (as described in *Materials and Methods*), we can ask whether geometrical symmetry underlies the low heterogeneity of these complexes. For example, a complex with precisely 4 copies of the same protein, $h_c = 1/4$, may be reliably assembled if the proteins are wedged so that they lock in a symmetric ring-like structure (Fig. 4*B*). The constraints that

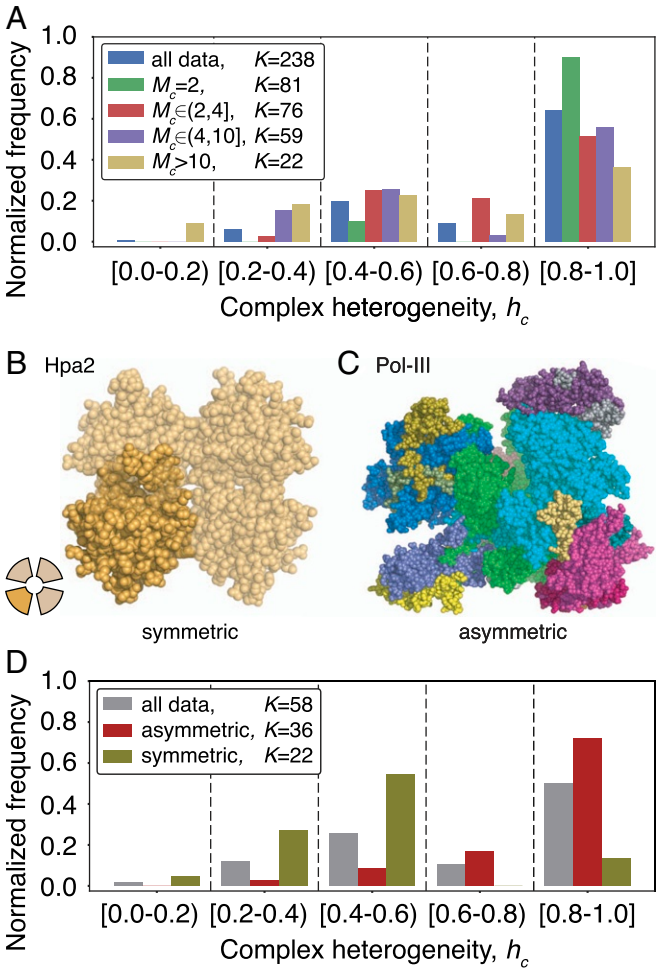

Fig. 4. Heterogeneity of symmetric and asymmetric complexes. (*A*) Histogram of complex heterogeneity using data from ref. 21. Most complexes are highly heterogeneous, but a small peak is also present for intermediate values of the heterogeneity. This distribution is preserved across complexes in different size ranges (different colors; see legend). (*B* and *C*) Structure of a complex with dihedral symmetry, Hpa2 (Protein Data Bank [PDB] ID code 1QSM), and an asymmetric complex, Pol-III (PDB ID code 5FJ9). The first has low heterogeneity, $h_c = 1/4$, whereas the second is completely heterogeneous, $h_c = 1$. (*D*) We cross-referenced data from ref. 21 with structural data taken from the PDB database (22). We then separated symmetric and asymmetric complexes. The high heterogeneity peak is only present for asymmetric complexes, and the peak at intermediate heterogeneities is only present for symmetric complexes.

limit assembly reliability in this type of complexes are likely to be different from those for asymmetric complexes (Fig. 4*C*). To estimate the role of symmetry in the heterogeneity of complexes, we classified them into those for which the crystal structure exhibits symmetry and those for which it does not (22). In Fig. 4*D*, we plot the corresponding heterogeneity histograms, which clearly show that asymmetric complexes have a very large heterogeneity bias, whereas symmetric complexes exhibit a large peak for intermediate heterogeneity values. We thus corroborate that high heterogeneity is indeed widespread among asymmetric complexes, to which we have limited our model. These conclusions should be taken with a grain of salt, however, since, generally speaking, the databases of protein complexes are in their infancy, and are prone to many possible methodological biases and ambiguities (see also discussion in *SI Appendix*, section G).

Discussion

Additional Mechanisms to Prevent Chimeric Assembly. Within the cell, assembly of complexes takes place in a dense mixture of proteins. At the basis of successful assembly lays the discrimination of the particular proteins of a complex among many others present in the mixture. Here, we argued that thermal physics puts strong constraints on the characteristics of complexes so that they assemble reliably. In particular, to keep protein promiscuity low, the heterogeneity and sparsity of complexes is constrained to high values. Clearly, within the cell, there are additional mechanisms that may help avoid chimeric assembly, some of which we describe now.

First, cellular protein concentrations can be spatially and temporally controlled to increase assembly precision and yield. In particular, cellular liquid droplets can provide local environments of high concentration of certain proteins, which may increase the assembly yield of corresponding complexes. A well-known example of cellular "compartmentalization" is the assembly of ribosomes inside the nucleolus, where ribosomal proteins are synthesized (23). Similarly, the temporal aspects of production and transport of proteins can be regulated to facilitate and optimize complex assembly (24). Second, it is highly plausible that the energy landscape for protein–protein interactions itself might have evolved to facilitate the kinetics of protein complex assembly. This is similar to the arguments given for other classical discrimination phenomena in biology, such as the discrimination of correct initiation DNA sites by transcription factors (25, 26). A third possible mechanism to prevent formation of chimeric structures is the usage of nonpairwise protein interactions, for example, allostery. For instance, it was suggested, in ref. 5, that different tetrameric receptor complexes of the bone morphogenic protein pathway assemble upon binding of particular ligands. Furthermore, because only 7 different species of proteins are involved in forming these receptors, the usage made of these proteins is dense, rather than sparse, which suggests that allostery provides a means to make a dense usage of proteins and enable combinatorial expansion. Finally, the geometry of complexes itself might have also evolved to optimize assembly. The presence of a peak at intermediate values of the heterogeneity in Fig. 4*A*, which can be ascribed to the symmetry of complexes, strongly suggests that the heterogeneity constraint that we derived may be avoided by means of geometric constraints to the structure of the complex, in line with the findings in refs 7 and 12.

Two organizational principles of cellular protein assembly explored in our theoretical model, namely the high heterogeneity of protein complexes and the sparse usage they make of the proteome, should be already functional at thermal equilibrium. So should be also the 4 additional mechanisms described above; therefore, they could be incorporated and studied within future extensions of our model. Going beyond equilibrium considerations, it is important to stress that the accuracy and the speed in protein assembly can be enhanced by a number of out-of-equilibrium mechanisms (27–29). For example, more than 200 nonribosomal proteins are involved in ribosomal biogenesis (30). Many of these are energy-consuming enzymes and have a variety of roles, for example, stabilizing protein–RNA interactions. At the same time, in vitro studies have shown that it is possible to assemble ribosomal complexes in the absence of such enzymes (9, 31, 32), albeit with a smaller yield. This evokes a possible analogy with the process of protein folding, which is typically facilitated and sped up by energy-consuming chaperones, but also can take place in their absence (33). It is tempting to propose that just as evolution had selected foldable proteins from the vast space of possible amino acid chain sequences, it might have also selected reliably self-assembling cellular complexes from the vast space of all possible multiprotein assemblies (Fig. 1).

Broad Distribution of Protein Participation in Complexes and Dynamic Control. An additional role of out-of-equilibrium processes is to control the dynamics of protein complexes. Indeed, many complexes are assembled in a contextual manner, that is, only when they carry out a function that is needed. Such highly dynamic complexes are often involved in regulatory and signaling functions (4, 5), and they can be contrasted with other more stable complexes (such as the ribosome), on which we have focused our attention here. The complexes including cyclins, which participate in the regulation of the cell cycle, provide an example of such highly dynamic complexes (6). The temporal sequence of assembly and disassembly events observed in this system is correlated with phosphorylation/dephosphorylation of the cell cycle components. Remarkably, the heat release in this out-of-equilibrium process has also been recently measured (34).

A possible footprint of regulated out-of-equilibrium phenomena is the existence of proteins that participate in many complexes, such as cyclin-dependent kinases, which have the potential to act as dynamic controllers. In order to assess the prevalence of such proteins, we analyzed several databases that contained information on protein participation in complexes (*SI Appendix*, section I). Our analysis suggests that the number of complexes, q_α, in which a given protein species α participates has a broad distribution, depicted in Fig. 5. Notably, this distribution cannot be simply explained by randomly grouping proteins into sets with the observed composition of complexes, unlike in other systems with shared components (35). The "excess" of highly participatory proteins can be viewed as an indication that this broad distribution might have evolved to carry out a particular function. One such function could be to assure an appropriate dynamic control of different assemblies (provided, of course, that this broad distribution is confirmed by future analysis of larger databases). A fascinating topic for future research should be, therefore, to extend our theoretical framework to allow dynamic, out-of-equilibrium control of complexes, and to verify whether the latter could indeed correlate with the observed excess of highly participatory proteins.

Protein Complexes as a Distinct Regime of Matter. Reliable self-assembly of protein complexes within the noisy cellular environment is intriguing not only from the point of view of cell biology but also from that of material science. Typical materials contain only a handful of different sorts of atoms, and the possible set of interactions between components is small. This results in states of matter that are easily reproducible: All crystals of salt are formed by Na and Cl atoms arranged in the same way. A different scenario is that of glassy materials, such as silicate glasses, in which the set of effective interactions among constituents is large due to spatial disorder. The large number of interactions makes the state of a glass unique and irreproducible: In each piece of glass, the arrangement of atoms is different.

Protein complexes combine similarities with both types of materials. Like glasses, they present a large number of effective interactions, although the origin of these interactions is not disorder but the large number of different specifically interacting components. As in crystals, the arrangements of these components may be highly reproducible: All ribosomes in a cell are made of many different proteins, yet the arrangement of the core ribosomal proteins is basically the same. This unusual combination of properties is rarely considered in physical theories of matter, with the closest analogue being "programmable materials" of biological origin, such as DNA origami (36) or self-assembling colloidal particles (37). One of the directions of future studies could be to further explore underlying principles of reliable equilibrium and nonequilibrium assembly for synthetic materials inspired by biology.

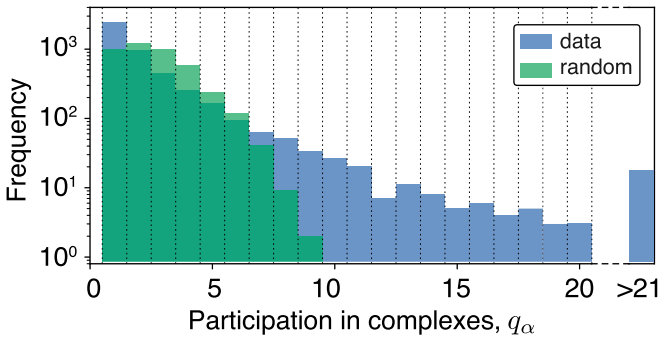

Fig. 5. Protein participation in complexes is not explained by null random model. In blue is the histogram for the participation of proteins in complexes (i.e., in how many complexes a given protein takes part) using the "core" dataset from ref. 40 (*SI Appendix*, section I). In green is the the same histogram for a dataset constructed by randomly grouping protein species into sets with the empirical composition of complexes. The null random model largely deviates from the data, and does not account for the prevalence of highly participatory proteins.

Materials and Methods

Model. Consider a set of N_{tot} component species labeled $\alpha = 1, \ldots, N_{\text{tot}}$, which form the "proteome" of our theory. By establishing "binding links" with each other, components can assemble into K different complexes labeled $c = 1, \ldots, K$. We index, by $\delta_\alpha = 1, \ldots, z_\alpha$, the binding links of a component of species α, where z_α is the valence of that species. How 2 components of different species, α and β, interact when their binding links δ_α and δ_β are in proximity is characterized by the binding energy tensor $U_{\alpha\beta\delta_\alpha\delta_\beta}$. If these 2 components are part of the same complex c in which they are bound to each other by linking δ_α to δ_β, they will interact strongly with an energy E, and so $U_{\alpha\beta\delta_\alpha\delta_\beta} = -E$. Conversely, if the components α and β are not bound together in any complex (or they are, but not by linking δ_α to δ_β), their interaction energy will be assumed null: $U_{\alpha\beta\delta_\alpha\delta_\beta} = 0$. However, even when the interaction energy between 2 components is null, these may still bind to each other through nonspecific interactions, provided that their concentration, p, namely, their chemical potential $\mu = \log(p)$, is large enough. One important quantity that characterizes the interactions of a component species α is its promiscuity, $\pi_\alpha = \sum_{\beta\delta_\alpha\delta_\beta} \Theta(-U_{\alpha\beta\delta_\alpha\delta_\beta})$, which is the total number of different species with which it has specific interactions. We also define the participation of the species, q_α, as the number of different complexes in which it takes part.

Note that we have made the simplifying assumptions that the strength of all interactions and the concentrations of all components are the same; these assumptions are relaxed in *SI Appendix*, section F. In addition, we will assume that the valence of all species is the same, $z_\alpha = z$. For the lattice implementation, we will further assume that the valence is given by the coordination number of the lattice, $z = 4$, while our analytical arguments are also valid for other values of z (*SI Appendix*, section C).

Although the structural or enzymatic characteristics of complexes largely define their function, here we are only interested in the characteristics that determine self-assembly. We quantify these characteristics through a small set of parameters. For each complex c, we define its *heterogeneity*, $h_c \equiv N_c/M_c$, as the ratio of the number of different component species in the complex, N_c, to the total number of components in the complex or complex size, $M_c \geq N_c$ (we have made the assumption that all complexes have the same number of components, relaxed in *SI Appendix*, section G). We also define the sparse usage that a complex makes of the available components, that is, its sparsity, as $a_c \equiv (N_{\text{tot}} - N_c)/N_{\text{tot}}$. For the lattice implementation, we have made the assumption that all complexes are perfect squares. This is an important simplification, which may particularly affect certain complexes, for which the assembly is strongly tied to their geometry (12). Note that the model in ref. 16 corresponds to the case $h_c = 1$ and $a_c = 0$ (in addition to fixing the ratio μ/E). Here, we relax these constraints, which allows us to explore the biologically relevant regime of protein complex assembly.

Figure Parameters. In Fig. 2, we considered $N_c = M_c = N_{\text{tot}} = 20^2$ and $K = 8$. The initial state corresponds to a fragment of the complex containing its central components (a nucleation seed). The fragment is of size 7×7, in

Fig. 2 *A* and *B*, and 20 × 20 (whole complex) in all other panels. In Fig. 2 *E* and *F*, each point reports the average of 3 replicate simulations (randomized complexes). Each simulation is run for a duration of 10^6 latt in a 40 × 40 lattice, with 1 latt corresponding to one lattice sweep. In Fig. 3 *B* and *C*, the parameters are as in Fig. 2, with $E = 7$, $\mu = -12.6$, and $K = 1$. In Fig. 3*D*, the parameters are as in Fig. 3 *B* and *C*, with $h_c = 1$ and variable K.

Data Availability. All data used in the paper were obtained from publicly available resources. See *SI Appendix*, Table S1 for a summary of the corresponding data sources. Details on the analysis are provided in *SI Appendix*, sections H and I.

ACKNOWLEDGMENTS. We thank David A. Huse and Joel Lebowitz for helpful discussions. We also thank Birgit H. M. Meldal for several exchanges regarding analysis of the data in ref. 24. This research has been partly supported by grants from the Simons Foundation to S.L. through the Rockefeller University (Grant 345430) and the Institute for Advanced Study (Grant 345801). P.S. is funded by the Eric and Wendy Schmidt Membership in Biology at the Institute for Advanced Study.

1. N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz, The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
2. H. Garcia-Seisdedos, J. A. Villegas, E. D. Levy, Infinite assembly of folded proteins in evolution, disease, and engineering. *Angew. Chem. Int. Ed.* **58**, 5514–5531 (2018).
3. S. Reuveni, M. Ehrenberg, J. Paulsson, Ribosomes are optimized for autocatalytic production. *Nature* **547**, 293–297 (2017).
4. A. Stein, R. A. Pache, P. Bernadó, M. Pons, P. Aloy, Dynamic interactions of proteins in complex networks: A more structured view. *FEBS J.* **276**, 5390–5405 (2009).
5. Y. E. Antebi *et al.*, Combinatorial signal perception in the BMP pathway. *Cell* **170**, 1184–1196 (2017).
6. A. W. Murray, Recycling the cell cycle: Cyclins revisited. *Cell* **116**, 221–234 (2004).
7. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244–247 (2017).
8. M. Johansson, J. Zhang, M. Ehrenberg, Genetic code translation displays a linear trade-off between efficiency and accuracy of trna selection. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 131–136 (2012).
9. A. M. Mulder *et al.*, Visualizing ribosome biogenesis: Parallel assembly pathways for the 30s subunit. *Science* **330**, 673–677 (2010).
10. A.-C. Gavin *et al.*, Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
11. E. D. Levy, J. B. Pereira-Leal, C. Chothia, S. A. Teichmann, 3D complex: A structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155 (2006).
12. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
13. J. R. Perilla *et al.*, Molecular dynamics simulations of large macromolecular complexes. *Curr. Opin. Struct. Biol.* **31**, 64–74 (2015).
14. J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer Associates, Sunderland, MA, 2001).
15. H. C. Berg, The rotary motor of bacterial flagella. *Annu. Rev. Biochem.* **72**, 19–54 (2003).
16. A. Murugan, Z. Zeravcic, M. P. Brenner, S. Leibler, Multifarious assembly mixtures: Systems allowing retrieval of diverse stored structures. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 54–59 (2015).
17. T. L. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics* (Dover, Mineola, NY, ed. 2, 1989).
18. C. P. Brangwynne *et al.*, Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
19. R. P. Sear, J. A. Cuesta, Instabilities in complex mixtures with a large number of components. *Phys. Rev. Lett.* **91**, 245701 (2003).
20. R. R. Kopito, Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol.* **10**, 524–530 (2000).
21. B. H. M. Meldal et al., Complex portal 2018: Extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2018).
22. H. M. Berman et al., The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
23. V. Sirri, S. Urcuqui-Inchima, P. Roussel, D. Hernandez-Verdun, Nucleolus: The fascinating nuclear body. *Histochem. Cell Biol.* **129**, 13–31 (2008).
24. S. Kalir *et al.*, Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292**, 2080–2083 (2001).
25. A. Tafvizi, L. A. Mirny, A. M. van Oijen, Dancing on DNA: Kinetic aspects of search processes on DNA. *ChemPhysChem* **12**, 1481–1489 (2011).
26. M. Cencini, S. Pigolotti, Energetic funnel facilitates facilitated diffusion. *Nucleic Acids Res.* **46**, 558–567 (2017).
27. J. J. Hopfield, Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4135–4139 (1974).
28. S. Pigolotti, P. Sartori, Protocols for copying and proofreading in template-assisted polymerization. *J. Stat. Phys.* **162**, 1167–1182 (2016).
29. G. Bisker, J. L. England, Nonequilibrium associative retrieval of multiple stored self-assembly targets. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10531–E10538 (2018).
30. D. Kressler, E. Hurt, J. Bassler, Driving ribosome assembly. *Biochim. Biophys. Acta Mol. Cell Res.* **1803**, 673–683 (2010).
31. S. Mizushima, M. Nomura, Assembly mapping of 30s ribosomal proteins from e. coli. *Nature* **226**, 1214–1218 (1970).
32. R. Röhl, K. H. Nierhaus, Assembly map of the large subunit (50s) of *Escherichia coli* ribosomes. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 729–733 (1982).
33. K. S. Hingorani, L. M. Gierasch, Comparing protein folding *in vitro* and *in vivo*: Foldability meets the fitness challenge. *Curr. Opin. Struct. Biol.* **24**, 81–90 (2014).
34. J. Rodenfels, K. M. Neugebauer, J. Howard, Heat oscillations driven by the embryonic cell cycle reveal the energetic costs of signaling. *Dev. Cell* **48**, 646–658 (2019).
35. A. Mazzolini, M. Gherardi, M. Caselle, M. C. Lagomarsino, M. Osella, Statistics of shared components in complex component systems. *Phys. Rev. X* **8**, 021023 (2018).
36. M. R. Jones, N. C. Seeman, C. A. Mirkin, Programmable materials and the nature of the DNA bond. *Science* **347**, 1260901 (2015).
37. Z. Zeravcic, V. N. Manoharan, M. P. Brenner, Colloquium: Toward living matter with colloidal particles. *Rev. Mod. Phys.* **89**, 031001 (2017).
38. S. O. Garbuzynskiy, D. N. Ivankov, N. S. Bogatyreva, A. V. Finkelstein, Golden triangle for folding rates of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 147–150 (2013).
39. R. Milo, R. Phillips, *Cell Biology by the Numbers* (Garland Science, New York, NY, 2015).
40. M. Giurgiu *et al.*, CORUM: The comprehensive resource of mammalian protein complexes–2019. *Nucleic Acids Res.* **47**, D559–D563 (2018).