


Modeling protein structure as a stable static equilibriumKaizhi Yue *Conformational Search Solutions, Palo Alto, California 94306, USA*

(Received 3 February 2022; revised 19 July 2022; accepted 24 July 2022; published 22 August 2022)

We present evidence that the protein structure can be modeled as a stable static equilibrium, determined mainly by compressive supports in the nonpolar interior. That is, protein structures derive their structural strength through the same mechanical principles as do conventional structures like bridges and buildings. This is based on the observation that the experimentally elucidated structural determinants, the interior nonpolar side chains, are engaged in strong compressions in static terms. At the same time, major substructures in proteins, helices and h-bonded strands, because of their geometry, inherently leave gaps in the space they occupy. Under the compressive force, nonpolar side chains from one substructure can protrude into the gaps of another neighboring substructure and block its motion. As a result, interlocking of substructures can form, which builds up the nonpolar core assembly. The native structure then is the one with the structurally most stable core assembly. While intuitively appealing, this is a radical departure from the prevailing thinking that protein native structure is determined by global energy minimum, which is founded on thermodynamic hypothesis. Furthermore, to develop an effective model for analyzing protein structure with conventional tools, a proper mechanical representation must be established. By proving that the stability of the equilibrium in compressive interactions is conditioned on a form of mechanical energy minimum, we show that our notion of native structure can be equally consistent with the thermodynamic hypothesis. By mathematically treating the blocking action, an interaction, as a bar, a physical object, we succeed in representing and analyzing the core assembly as truss, a conventional structure. In this paper we define and expound step-by-step increasingly integrated interlocking patterns. We then analyze the core assemblies of a large set of diverse protein database structures. A native structure can be distinguished from decoys by comparing the composition and strength of their core assemblies. We show the results for two sets of native structures vs decoys.

DOI: [10.1103/PhysRevE.106.024410](https://doi.org/10.1103/PhysRevE.106.024410)**I. INTRODUCTION**

There have long been suggestions and conjectures that proteins must have assumed their structures on the basis of simple mechanical principles. These usually arise from observing the simple and elegant geometry of proteins. Richardson [1,2], Chothia [3], Salemme [4], and many others [5] have given comprehensive and in-depth discussions on geometric relations governing the organizations of protein structures, often alluding to the necessary statical causes or effects. In particular, a large body of literature describes the *knobs into holes* and *ridge into groove* models of helical packing ([6] and references therein). Also extensively explored are the twists of β strands, the twists in intrasheet strand packing [3,7,8], tight turns [1,2], helical packing in 4-helix bundle proteins [9], coiling of β hairpins, packing of α helices onto β -pleated sheets [10], packing of α/β barrels [11], and of β sandwiches [12], residue pairing in antiparallel strand packing [13], $\beta\alpha\beta$ turns [14], and side chain organization in type I tight turns [1]. These pioneering investigations have laid the foundation for identifying and elaborating the underlying statical interactions in protein structure.

In the meantime, experimental and theoretical research in the process of protein folding has demonstrated unequivocally that the driving force is the collapse of the nonpolar portion of the protein chain in the polar solvent, water. After Tanford

[15] demonstrated in experiments, many authors forcefully argued for it theoretically (see [16] and references therein and [17–23]). Energy potential functions have been developed for characterizing protein folding process, some specifically for accommodating solvent effect, e.g., EEF1 and others [24–33].

The observations on the mechanical origin of protein structures are further strengthened by experiments on homolog and mutated proteins [34–40]. It is shown that even when sequence identity is as low as 40%–50%, as long as the same *key residues*, mostly large nonpolar residues, are preserved, the protein structures can be virtually intact. Essentially, these extraordinary results suggest that despite the large number of atoms, atomic groups and their myriad of interactions, to develop a model that can account for the three-dimensional structure and associated stability of proteins, one needs to focus only on those *key residues*. To wit, there can be a model that is simple yet effective. More recently, research on high pressure response of mutant structures has shown that there seems to be considerable rigidity in protein interior and it should arise from the side chain (SC) interactions [41].

Our research has also benefited from the work in the field of structure prediction. To verify our hypothesis, we must compare native structures and non-native ones. High quality decoys are indispensable. In particular we have taken advantage of the databases of decoys from various labs [42–45].

A. The need and the challenges in establishing a statics-based model

Our research suggests that the simple mechanical principles that determine the protein structure can be basically just statics. This means, to some extent, the protein structure can be viewed just as everyday structures like bridges and buildings. This then implies we will be treating protein structure as an equilibrium of forces and the native structure will be the one that can withstand the strongest mechanical impact. While sounding intuitive and innocuous, as a thoroughly developed model, this is a radical departure from an established principle in theoretical and computational protein research, the search for *global energy minimum* [46].

In past decades, methods of molecular mechanics (MM) and molecular dynamics have been developed for investigating stability of biomolecules including proteins. These have contributed enormously to our understanding of biomolecules. In applying these methods to protein structure, the widely accepted tenet postulates that its stability is based on the energy minimum of some energy potential [46]. This potential is often taken from a force field of MM and at atomic level [24,47–50]. Alternatively, it can be an elaborately designed lower resolution potential [51–55]. This tenet is in turn based on thermodynamic hypothesis, which states that the protein native structure must be an accessible minimum energy state. The potential for the protein-solvent system is often treated as a sum of all the component potentials including bonded potentials, such as bond stretching and bond angle bending, and nonbonded potentials, such as electrostatic (ES) and Lennard-Jones (LJ) potentials [46]. The energy minimum of the system will be the minimum of that scalar sum. Thus, seeking global energy minimum has become a principle for investigating protein structures.

This principle of global energy minimum has guided theoretical and computational research on protein stability and been considerably fruitful. But, it is not without shortcomings, especially with identifying native structures. Koehl and Levitt have famously lamented about “a central embarrassment of molecular mechanics, namely, that energy minimization or molecular dynamics generally leads to a model that is less like the experimental structure” [56,57].

It is tempting then to consider a static model. Not only we have aforementioned evidence supporting a notion of stability based on force balance, but, also, it is conceivable that in a simple mechanical construction a scalar sum may not be an appropriate indicator of structural strength.¹

Yet, until now there is no serious attempt on an alternative, e.g., a statics-based model, because there are formidable difficulties. First, the thermodynamic hypothesis dictates that the protein stability must be associated with an energy minimum.

¹To illustrate the difference between seeking minimum of total energy and seeking structural stability (see Sec. III), we can refer to Fig. 3. Imagine we replace all the line segments by springs and these springs are at their equilibrium position. The strain energy will be zero. Imagine we now move spring *AD* to coincide with *AC*. Clearly, the total strain energy is again zero. But, we know the former is determinate and much more stable.

Even if a static model can formulate a perfect equilibrium for the protein-solvent system, unless it can successfully establish some form of energy minimum, it is theoretically suspect in its soundness. In particular, to assume that the protein structure is determined by statics is *often perceived as* to assume that there exists a certain gadgetry that “locks” its parts into a fixed shape. This to some extent goes against thermodynamic hypothesis. Thus, it is rarely contemplated.² Second, it is hard to conceive a structure whose underlying support comes from compression, let alone a molecular structure. In the first place, compression is repulsion in energy potential terms which is often associated with destabilizing structure. Furthermore, it is difficult to identify the force which can counter the compression in the interior to build an equilibrium. Third, to develop an effective model for analyzing protein structure with conventional tools, ways for representing the core assembly as a conventional structure must be devised. In particular, a proper way of modeling compression must be found. We will show that all these difficulties have been handled in our model.

B. Relevant concepts in traditional structure research

Traditional disciplines such as structural mechanics, mechanics of material, and structural analysis have been developed for analyzing those conventional structures on the basis of principles of theoretical mechanics, mostly statics.

The methodologies and conclusions of traditional structure research that are relevant to protein structure can be summarized as follows: Interactions between *structural members*, e.g., stone, brick, cable link, column, and beam, determine the shape or topology of a structure. The stability of a structure depends both on the building materials and the way the structural members are arranged. The structural stability is analyzed through equilibrium of forces and moments in three dimensions (3D). A structure may fail because a structural member breaks when reaching the strength limit of its material. A structure may also fail in the form of *buckling* [61–65]. That is when no structural member is broken but the equilibrium state shifts onto an unsustainable path. This is worth mentioning because protein unfolding could be of this mode of structural failure. An example of buckling is when one pushes two ends of a slender stick (see Fig. 1). Supposedly, if the pushing is perfectly centered, the stick will only be compressed. But, in reality inevitably when the force exceeds a threshold, the stick will bend, and eventually break. This threshold is the famous Euler’s *buckling load* (alternatively termed “critical load”).

The term *load* here refers to the various external force actions everyday macroscopic structures are subject to. These include wind, earthquake, etc., but the most common and ubiquitous is gravity. The load level at which the structure

²The only known model that can be considered mechanical is the cardboard box model [58]. It is a model on the folding process (rather than on protein stability like ours) and is only qualitative. It does not have significant followup research, which may have to do with the fact that the supporting experimental data were later challenged [59,60].

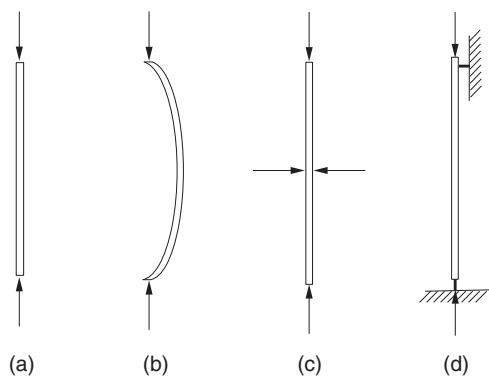


FIG. 1. Equilibrium states and stabilization of a stick or *column* in engineering terms. In (a) a column is being pushed. In (b) under a large enough compressive load (i.e., near the buckling load) a straight slender column will bend and eventually break. (c), (d) Show the additional supports in the middle (c) and at the ends (d). These will increase the buckling load fourfold or twofold, respectively. More detail is given in Sec. VI A 2.

fails is called *failure load*. To counter the load and support the structure, structural members are designed to carry and distribute the load. Usually and ultimately the *support* comes down to the ground. In this paper, we will show that there are counterparts of these in the protein structure but the similarity will be at a very general level. For example, the load in the context of protein structure would be a perturbation, such as a thermal impact.

C. Specific static properties of protein structure

The protein structure, arising from nature, is fundamentally different from manmade structures. Three of its specific properties are of central importance: the characteristics and source of the supporting force of the structure, the mechanisms for fixating its shape or topology, and the mechanism for keeping the equilibrium stable. We briefly explain these in turn.

Although the protein structure is in a dynamic equilibrium, it can be argued that there is no fundamental difficulty in applying statics to the analysis of a protein database (PDB) structure. This is because the structure in 3D coordinates is a time and ensemble average anyway. The real problem is exactly what is the source of the force that consolidates the protein structure and what is the mechanism that determines the topology of the structure. If we examine the interactions involving the deeply buried nonpolar SCs, these are uniformly repulsions, i.e., behaving essentially as *compressions* in structural analysis terms. When thinking in terms of repulsions, the forces may be perceived as destabilizing. But, as compressions they are widely used for structural supports. Arches, domes, and buttresses in conventional structures are well-known examples. Then where do these compressions come from? The fact that a protein chain collapses from an extended state to molten globule state and further condenses to the folded state provides the answer: The entropic inward normal force from hydrophobic effect is the main source of the compression in the protein interior.

Not only the interior nonpolar SCs are engaged in compressions, the SCs in one secondary structure can protrude into the

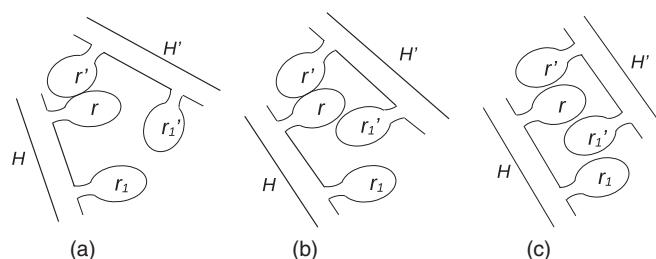


FIG. 2. Schematic illustration of blocking (a), double blocking (b), and interlocking (c).

gap between SCs of another, blocking the axial motion of the latter. The helices and h-bonded strands in protein structure, which we will refer generically as *substructures*, have regular geometry. In the regular arrangement of these substructures, sufficiently large gaps between sequential neighbors will form. As such, the intrusion by a SC from another substructure has a sufficiently high probability. When the probability is too high, the intrusions and blockings will become random and that may cause protein chain to assume multiple topological forms and there is no stable unique structure. But an analysis of the data shows (see Supplemental Material [66]) that there may be just enough such gaps so that if there are opportunely positioned large SCs to intrude or protrude into them, then we will have the key residues that determine a unique structure for the chain.

Specifically, the axial interactions between nonpolar SCs from two substructures can grow stronger with concomitant multiple forces. It is easy to infer possible combinations. To name a few, a nonpolar SC r' from substructure H' may protrude into a gap between r and another SC of substructure H and has a strong axial repulsion with H at SC r . We refer to such a case as *r' blocking r* [Fig. 2(a)]. If there is a sequential neighbor of r' , r'_1 , also from H' that blocks r from the opposite direction, then the axial motion of r and to a considerable extent the axial motion of H will be constrained. We refer to such a *double-blocking* case as *locking* for short. In Fig. 2(b), r' and r'_1 lock H at r . Two substructures can mutually double block or lock, then we have *interlocking* [Fig. 2(c)]. Conceivably, there can be various combinations of blockings, lockings, and interlockings. The interior core of the protein can be assembled this way. Upon this assembly the full protein structure can be built. Compressive supports are by nature strong but unstable. The aforementioned column buckling is a prominent example. Stabilization mechanisms are often necessary for structural members like column. For proteins, the blocking action must be stabilized and the inherent stabilizing environment is the multitude of nonpolar contacts in the interior. An equation for the buckling load of the blocking action is derived in Sec. VI.

Our survey of a large set of PDB structures shows that every structure has an *assembly* of substructures whose topology is fixed by the above mentioned blockings in complex blocking patterns. In contrast, for the decoys that are tested, most do not have as strong an assembly. Some decoys, in particular decoys produced through threading, do have a similar assembly. But they usually have buried charged or strong polar SCs which will destabilize the assembly.

This paper is organized as follows: We first introduce geometric-mechanical patterns of (mostly) SC interactions, e.g., blocking or interlocking, at increasing levels of organizations, until the level of nonpolar core assembly. To mathematically treat core assemblies, we represent them as a simple conventional structure, the truss. This will allow us to solve the equilibrium equations of a core assembly and analyze its structural strength, in particular its load distribution and structural determinacy. We then consider comparing stabilities of these core assemblies on the basis of the structural strength. Finally, we present the results in applying our analysis to native structures and decoys. In the Discussion section, we detail the considerations for some modeling decisions, in particular the buckling load of a blocking action.

II. MECHANISMS OF INTERLOCKING

A. Molecular forces in a static model of protein structure

Static analysis means both the whole structure and each body in the structure must be in an equilibrium state. To wit, forces and moments must be balanced in three translational and three rotational degrees of freedom (DOF). Now we introduce the bodies and forces considered in our model.

1. Solid bodies in protein structure

The core components of protein structure can be approximated as *solid bodies*. Here we will be using the term in its classical interpretation: a solid body is a mechanical body that is *not necessarily rigid* but has a definite shape. Within a range it is able to recover its original shape after a deformation. In the scope of this presentation, the solid bodies are mainly spontaneously formed helices and h-bonded strand pairs. Both these are secondary structures and can be modeled as cylinders. When we refer to substructure axis, it is the axis of the cylinder, which can be readily calculated from PDB files [12,67].

In this presentation, we adopt a coarse-grained residue level representation when analyzing the equilibrium of the structure. However, when obtaining the interaction forces, the calculation is at a more refined level. Each SC or main chain (MC) unit can be explicitly positioned, through their centroids, on the body according to the PDB coordinates.

All the coordinates are still given at atomic level using PDB files. All the force field (FF) level forces can still be calculated at atomic levels. But, only the resultant forces at SC or MC unit levels are used in analysis. The term *resultant* here refers to the sum total of the atom-atom forces between two atomic groups.

2. Forces

All the major forces considered in our static model are taken directly from the FFs widely used in the community, in particular, CHARMM, Amber, and Gromos [47–49]. In this presentation, we will use mostly Gromos FFs and adopt its standard *force unit* of $\text{kJ mol}^{-1} \text{nm}^{-1}$ [49]. We are concerned mainly with the determinant of relative positions of solid bodies. Thus only nonbonded forces are of interest. Furthermore, we consider strong forces only. Thus, we consider short range forces only for LJ interactions. For ES forces, we only con-

sider interactions between charged, strong polar, or MC polar groups and only when they are either in the protein interior (where the relative permittivity is low) or are in close distance (so that the shielding effect of water is low). MC h-bonds for helices are ignored as they are internal to the solid body and do not participate in interactions between substructures.

For the forces defined in standard FFs, only the above-mentioned nonbonded forces, LJ and ES, are explicitly considered. Salt bridges (SB), i.e., interactions involving charged groups, and h-bonds (HB), i.e., those between polar groups, are handled through combining ES and LJ forces [46]. Bonded forces are used in modeling the elasticity of MC-SC links but do not appear in the model's system of equilibrium equations per se. Whether a force is tension or compression is referred to as its *sense*. Whether a structural member or mechanism is capable of either or both, a spring or a steel bar being an example of the latter, is critical in how it contributes to the equilibrium of a structure. As we have argued so far, a compressive force may not be destabilizing for the whole structure. Yet, often a tensile force acts steadily whereas a compressive force unsteadily as we have pointed out. LJ forces, when they are at non-negligible level, are always compressive. ES interactions can be both compressive or tensile. The sense of a resultant force is based on the direction of the vector that is the sum of all the atomic forces.

We also consider π stacking [68] between aromatic rings. The force magnitude is estimated through the average interaction energy and interacting distance. The direction is between the centroids. The sense is attractive. At the same time, the two aromatic rings are often engaged in strong repulsion. But, the coexistence of attraction and repulsion does not imply a conflict (or a cancellation if a scalar sum is being calculated). First, their effects peak at different distances. When they are very close, the attraction part relents, i.e., becomes relatively negligible, then when they are at a distance the attraction takes over. Second, they can serve the same purpose. A way of viewing the attractions and repulsions in π stacking is that it is linking an arch's stone wedges with powerful springs, thus doubly stabilizing the structure.

We consider two nonpairwise forces as well. First, it is the *entropic* hydrophobic force. This force can be viewed as associated with the desolvation free energy of a nonpolar group or associated with the notion of solvent exposed surface area [19,21,46]. If we consider that hydrophobic effect is to minimize the exposed nonpolar surface area of a protein, and if we also consider the force field conservative, i.e., we can have a potential, then differentiating the potential we should mathematically have a force. This is a force acted on a nonpolar group, pointed inward and basically normal to the surface. The hydrophobic effect is the most fundamental to protein structure. In that sense, the hydrophobic force provides the ultimate support to protein structure just like ground support does for conventional structures. This force will show up in the equilibrium equations as *support reactions*, the external forces that counter the load.

Second, we consider the neutralizing force for a charged or strong polar group that is buried in the nonpolar interior. When it is not neutralized by forming a SB or HB with a close-by charged or polar group, such a group will be subjected to

a very strong outward force. Unlike the hydrophobic force, this force can be derived from a regular FF by considering the electric field of the standalone charged or strong polar group.

For both forces, their directions are based on geometry. In particular, they depend on properly defined nonpolar interior. This “interior” notion is related to but different from solvent accessible surface as defined in [69]. A nonpolar interior can contain a large enough void for some water molecules to move in. But, the relative permittivity can still be different from that of bulk solvent. This is discussed in detail in Sec. IV D.

B. Blocking, double blocking, and mutual blocking

We now formally define and expound the basis of our model, the mechanisms of interlocking. As explained in the Supplemental Material [66], the axial direction displacement of the substructures is assumed to be the most common scenario for a structure to deform and disintegrate. Thus, we start from the action that blocks this motion.

We first define blocking of a substructure H at SC r by a SC r' from a substructure H' as shown in Fig. 2(a). For this, we consider the resultant LJ force $f_{rr'}$ between SCs r and r' . If we denote the axis of H by \mathbf{a}_1 , then the axial component of the force will be $f_{rr'} \cdot \mathbf{a}_1$.³ There is a *blocking* f_b if this axial component exceeds a threshold $\delta > 0$, i.e.,

$$|f_b| = |f_{rr'} \cdot \mathbf{a}_1| \geq \delta. \quad (1)$$

If a blocking is strong enough in this axial dimension, it can stop the motion of the substructure H in one direction, thus contributing to the reduction of DOF of H . The intensity with which this axial blocking acts is naturally the axial component of $f_{rr'}$, f_b . For notational convenience later, we can introduce a relation $\text{block}(\{H, r, \pm\}, \{H', r', \pm\})$ to mean Eq. (1) where “ \pm ” is “+” or “−” indicating the sense of blocking relative to the axis of the substructure H .

We emphasize that this is not the same as the *structural strength* of this blocking. The latter would mean at which force level the blocking will fail and the reduction of DOF will be reversed. It could be much larger. The failure of this blocking is likely closer related to the force $f_{rr'}$ than its axial component f_b . We thus consider this strength as $f_{rr', \max}$. The accurate measure of this strength requires detailed analysis of the nature of the geometry and force field involved in this LJ interaction between a SC pair. For now we would simply take the former as proportional to the latter. That is, approximately

$$f_{rr', \max} \propto f_{rr'}. \quad (2)$$

where “ \propto ” denotes *proportional*. If $f_{rr'}$ is in a near linear region of the LJ force (it can be so approximated as indicated in Sec. VI), we can introduce a notion of stiffness for such a force so that within a range

$$f_{rr'} = k_0 \Delta_0, \quad (3)$$

where k_0 is the stiffness and Δ_0 the displacement, i.e., the distance change between the interacting pair in the LJ force. Then we can further introduce a notion of stiffness of blocking

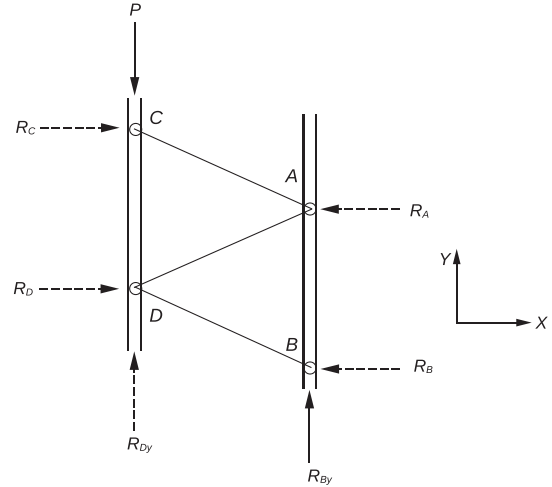


FIG. 3. An example of four points in equilibrium. The line segments connecting them, including double lines (between the points), represent the internal forces. A double line represents MC connection that can carry both compression and tension. A single line indicates a blocking. P and R 's are external forces. P can be taken as the load. R 's are support reactions. This setup can also be viewed as a simple truss with four joints and five bars (see Sec. III). The circled points are joints. The lines or double lines are bars.

k , $f_b = k\Delta$, where $\Delta = \Delta_0 \cdot \mathbf{a}_1$ is the axial projection of Δ_0 . Clearly, $k = k_0$. The reader may notice that we are making a big leap here: we are assigning stiffness to an interaction rather than a physical object. However, as will be seen in Sec. III, only with this notion of stiffness in place, we can solve structures using strain energy-based method in traditional structural analysis.

This blocking action exists in the context of a *maintained equilibrium*. For example, at the points of action, SCs r and r' , for the two SCs to be in equilibrium, $f_{rr'}$ must be balanced by some other forces. For this purpose, each SC can easily get some support from its respective MC unit. But, supports from lateral directions are also needed. We show a textbook example of such force equilibrium in Fig. 3. Here points A, B, C, D can be viewed as SCs attached to MC, while AB and CD represent stretches of MC. For example, f_{AC} can be like $f_{rr'}$. One may notice that there is a parallel between this figure and Fig. 2(c). Indeed, the former can be a formal representation of the latter. In this setup, for C to stay in equilibrium, if $f_{AC} = 0$, then the external load P can be balanced by f_{CD} . However, if we want to add f_{AC} to share the load with CD , another force must be added: the *support reaction* R_C at C . The equilibrium can also be affected by the asymmetry in force senses of the blocking action. For example, if AD can only carry compression as in a blocking, then in $+x$ direction, there will be a nonzero force (if f_{AC} is nonzero). There will need a reaction force R_A at A to balance it. As we will see in Sec. III this asymmetry will have severe consequences in developing our model.

Aside from the equilibrium per se, there is another factor in structural stability: *the equilibrium must be stable* [70,71]. The force $f_{rr'}$, being a compressive force, will be unsteady if the two acting bodies, SCs r and r' , do not receive some support actions, usually from lateral directions. In Sec. VI B

³In this paper, a vector will be in boldface if it is operated on as a vector.

we will show that a blocking action laterally supported by a single SC that performs like a spring will have a buckling load (i.e., above which the two SCs will slip off one another)

$$P = Ckr^2/L, \quad (4)$$

where C is a constant, k is the stiffness of the spring, and r and L are geometric parameters for the supporting and supported SCs, respectively. In the interior of a well-packed structure, the packing density is large enough. Each SC is allowed to have several neighbors from different directions. Thus, the force $f_{rr'}$ can be held steady by them.

1. Double blocking

If there are blockings in both directions of a substructure, its degrees of freedom in the axial dimension are removed or diminished, as shown in Fig. 2(b). Formally, substructure H is axially blocked both ways at SCs r, r_1 , by SCs r', r'_1 from substructure H' when

$$\begin{cases} |f_{rr'} \cdot \mathbf{a}_1| \geq \delta, \\ |f_{r_1r'_1} \cdot \mathbf{a}_1| \geq \delta, \\ (f_{rr'} \cdot \mathbf{a}_1)(f_{r_1r'_1} \cdot \mathbf{a}_1) < 0. \end{cases} \quad (5)$$

When the same SC is subjected to blocking in both directions, i.e., when $r = r_1$, there will be little if any moment. Furthermore, the efficiency of the use of SCs is improved. This pattern is frequently observed in protein structure; we thus define it as *locking* or *double blocking*, denoted by a relation $\text{lock}(r, \{r', r'_1\})$.

A locking stabilizes its blockings. By examining the equilibrium condition of the three interacting SCs, we can see why. Here not only the forces $f_{rr'}$, $f_{r_1r'_1}$ have axial projections that can hold SC r in place, but also $f_{rr'}$, $f_{r_1r'_1}$ can each prevent the other from slipping off the action line.

2. Mutual blocking

We observe that the formulation of Eq. (5) has a symmetric case in which f and a trade places:

$$\mathbf{f} \cdot \mathbf{a}_1, -\mathbf{f} \cdot \mathbf{a}_2 \geq \delta. \quad (6)$$

Here we assume force $-\mathbf{f}$ is on r' of substructure H' and is along the axis \mathbf{a}_2 . If it is opposite \mathbf{a}_2 , a minus sign “ $-$ ” must be added in front of \mathbf{a}_2 . These equations describe the case of *mutual blocking*,

When both double and mutual blockings happen, we will have interlocking, the main subject of the next subsection.

C. Interlocking between two substructures

As mentioned earlier, depending on the force projections on the axes of two interacting substructures, the forces may or may not generate blocking on both substructures and in both directions. However, if they do, the effect on fixating the structure is multiplied. An example of this setup is shown in Fig. 2(c). This relation can be formally specified in terms of four blockings on both substructures and in both directions. It will involve total eight SCs, $r, r_1, r_2, r_3 \in H$ and $r', r'_1, r'_2, r'_3 \in H'$.

Just like in the introduction of double blocking or locking, this general case is inferior in many properties, e.g., efficiency

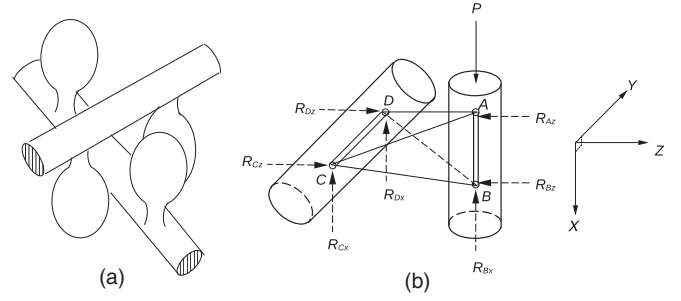


FIG. 4. Cross interlocking shown in schematic (a) and simplified (b) drawings. Note that the two cylinders in (b) represent the two substructures. The line segments AC, AD, BC, BD represent the blockings. AB, CD represent the MC parts. (b) Can be viewed as a truss representation which is detailed in Sec. III.

of SC usage and steadiness from the same SC being blocked. If not only r and r' are mutually blocked, but also r vs r'_1 and r' vs r_1 , then the number of SCs involved will change from eight to four. This pattern of blockings, with its advantages in properties, is also frequently observed. For example, it occurs when two helices are in near parallel packing and the blockings are arranged in tandem. Sometimes there are one or two blockings missing due to the force being slightly lower than the minimal intensity (δ). But, the pattern is unmistakable. We thus define this as *tandem interlocking* and refer it as *interlocking* when no confusion will arise. To express the case in terms of locking and blocking:

$$\begin{aligned} &\text{interlock}_{\text{tandem}}(r, r_1, r', r'_1) \\ &\equiv \text{lock}(r, \{r', r'_1\}) \text{ and } \text{block}(\{H, r_1, \pm\}, \{H', r'\}) \text{ and} \\ &\quad \text{lock}(r', \{r, r_1\}) \text{ and } \text{block}(\{H', r'_1, \mp\}, \{H, r\}), \end{aligned} \quad (7)$$

where \pm, \mp can be either “ $+$ ” or “ $-$,” $r, r_1 \in H, r', r'_1 \in H'$. When only the substructures are concerned, we may write $I(H, H')$ as a shorthand and may add subscript to “ I ” to indicate the type of an interlocking. By definition, $I(H, H') = I(H', H)$. Similar to how the strength of locking is defined, the strength of interlocking is just those of lockings separately on two substructures. We stress that “interlock” and “ I ” here should be interpreted as a shorthand for the physical equations implied by the “lock” and “block” relations, which we have commented when introducing the relation “lock.”

1. Cross interlocking

Two substructures can interlock in such a way that each has only two SCs involved yet both SCs are in locking. We call this cross interlocking. Schematic drawings are shown in Fig. 4. Formally this is equivalent to

$$\begin{aligned} &\text{interlock}_{\text{cross}}(r, r_1, r', r'_1) \\ &\equiv \text{lock}(r, \{r', r'_1\}) \text{ and } \text{lock}(r_1, \{r', r'_1\}) \text{ and} \\ &\quad \text{lock}(r', \{r, r_1\}) \text{ and } \text{lock}(r'_1, \{r, r_1\}). \end{aligned} \quad (8)$$

An example of cross interlocking in PDB structure 1ARQ is shown in the Supplemental Material [66].

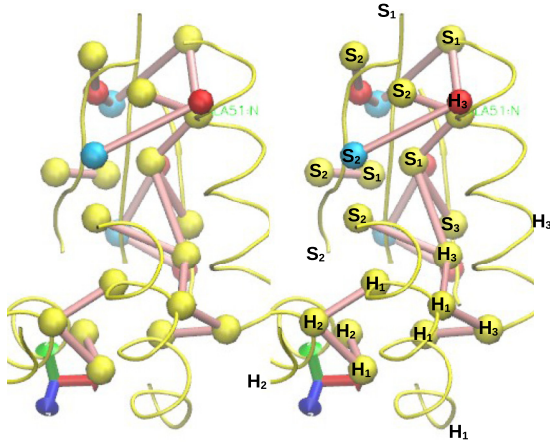


FIG. 5. A simplified display of the core assembly of ICTF, shown as stereoscopic images rendered with VMD [73]. Only non-polar and salt-bridge interlockings are shown as pink lines. These should be lines in the direction of resultant forces between two SCs. Because the large size of SCs will block the view, they are approximated by connecting close distance atom pairs from two SCs. For nonpolar interlocking, the atoms are carbon (colored yellow). For salt-bridge interlocking, they are oxygen (red) or nitrogen (blue) atoms. To show which substructure a SC is from, we have marked the SCs by “H1,” “S1,” etc. where “H” indicates helix and “S” strand. To make more SCs visible, their size is reduced. Note that since all the blockings are mutual, the pink lines for nonpolar interlockings are often nearly parallel to the two substructures from which the SCs emerge. This can also be viewed as a truss representation of the assembly (see Sec. III). In a truss representation, there should be bars for connecting joints of the same substructure. Here they are omitted to avoid cluttering the view. Instead, the cartoon images of substructures are added.

2. Hairpin, salt bridge, π stacking, and disulphide bond-based interlocking

Attractive actions such as h-bonds, salt bridges [72], π stacking, and disulphide bonds can also form interlockings. In addition, LJ forces can participate in an interlocking that is based on a particular substructure position, the hairpin.

The hairpin configuration is well known [1]. In our model, in the implementation, it is formally required that a hairpin is two substructures, helices or strands, connected by a loop with at most three residues in a near parallel position. The two ends of each of the substructures can be distinguished as connected end and opposite end. Then, a hairpin-based interlocking is defined as a hairpin that contains two blockings, one each on the substructures and pointing to the opposite end.

Similar to a hairpin, SBs can also reduce the DOF of a pair of substructures. We consider that there is an interlocking if there is a SB force at 180 force units (i.e., $\text{kJ mol}^{-1} \text{nm}^{-1}$) or there are two SB forces each at above 90 force units and at the same time there is a nonpolar locking between the two. (Examples of SB interlocking are shown in Fig. 5 and in 1AEP display in the Supplemental Material [66].) Since π stacking involves two aromatic rings, π -stacking-based interlocking is subsumed by nonpolar interlocking, only with higher strength. In particular, its attraction strengthens that

interlocking. Lastly, a disulphide bond alone can introduce an interlocking if it is between two substructures.

D. Assembly of substructures on the basis of interlocking

When we have more than one pair of interlocked substructures, we may investigate if they are fully fixated in space and, if so, investigate the property of the integrated body of substructures. We refer to any such collective body as a core assembly. If we use the concise notation introduced earlier, an example assembly containing substructures A, B, C, D may be specified as

$$I(A, B) \text{ and } I(B, C) \text{ and } I(C, D) \text{ and } I(D, A). \quad (9)$$

The above example can be generalized to define a core assembly as

$$\prod_{i,j \in \{(i_1, j_1)\}} I(H_i, H_j), \quad (10)$$

where Π denotes repeated “and” operation and $\{(i_1, j_1)\}$, a set of integer pairs. $\{(i_1, j_1)\}$ has two properties: (1) it is a subset of $\{(i, j) | i, j = 1 \dots m \text{ and } i \neq j\}$, assuming there are m substructures in the conformation; (2) the pairs of $\{(i_1, j_1)\}$ are in a cluster. That is, they are all connected.⁴ For convenience, we may also refer an assembly as the set of its constituent substructures $\{H_i\}$ or the set of its interlockings $\{I(H_i, H_j)\}$ when no confusion will arise.

We have commented at the time when the relationship “ I ” is introduced: it represents full 3D interactions among bodies. Here, the above relation, Eq. (10), specifies a full 3D physical assembly. We can infer properties of an assembly at this abstraction level. But, we can also flesh the assembly out to full 3D coordinates and carry out calculations. This would be solving the system of equilibrium equations for the assembly. In the very next section, we introduce the truss representation that offers a simplified framework for setting up this equation system. We show the stereoscopic displays of the core assembly of ICTF in a simplified representation in Fig. 5. A similar display for 1AEP is in the Supplemental Material [66].

There are a myriad of variations in configuring core assemblies from interlockings. These often correspond to packing configurations and some have been cataloged by the pioneering analysis work on protein geometric properties cited in the Introduction. What makes our treatment different is analyzing the same patterns using statics to examine the consequences in terms of force equilibrium and thus a well-defined structural strength and stability. As an example, we describe two common interlocking patterns in core assembly, high coordination and circularity. They will be formally defined here and analyzed in Sec. IV.

⁴This specific clustering is as follows: Let $A = \{i\}$ be a set of integers, $S_I = \{(i, j) | i, j \in A\}$ a set of pairs of integers in A ,

for all $i, j \in A$, there is a sequence σ_I

$$= [(i_k, i_{k+1})]_{k=0}^{l-1} \text{ such that}$$

$$[(\text{for all } (i_k, i_{k+1}) \in \sigma_I, (i_k, i_{k+1}) \in S_I)$$

$$\text{and } (i_0 = i, i_l = j)].$$

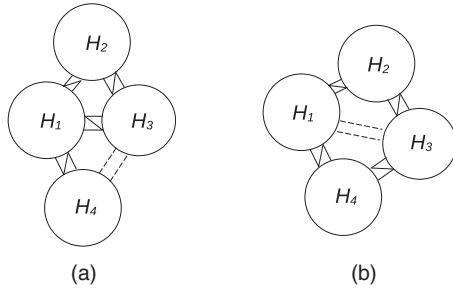


FIG. 6. Core assemblies with high coordination interlocking and circular interlocking. The connecting lines indicate interlocking. H_i 's are nearly parallel-packed helices. An interlocking can be added to each assembly so that the two have identical interlockings (Sec. IV B). The dashed lines indicate the additional interlocking. Then these are core assemblies of a 4-helix bundle.

High coordination configuration. A substructure can be chosen out of several as the center and the other substructures interlock onto it. As a result, the center one is a *high coordination* substructure. Again using Π to denote repeated “and” operations, we have

$$\Pi_{k=1\dots m} I(H, H_{i_k}),$$

where $\{i_k\}$ is a subset of $\{i | i = 1 \dots n\}$, n being the total number of substructures. That is, substructure H interlocks with all m substructures. m is the *coordination number* denoted by $C_H = m$. Clearly, the average of C_{H_i} is an indication of how interlocked the substructures are in the assembly. An example of this is shown in Fig. 6(a).

Circular configuration. A substructure usually can not interlock with too many neighbors. Thus, an efficient arrangement may have two high coordination substructures. But, then the interlocking between the two can become most responsible for the loss of the interlocked substructures. In parallel packing configuration, interlocking of substructures can form *circles*. This strengthens the assembly since breaking of one interlocking, which can be considered *redundant*, will not cause the assembly to lose a single substructure. For example, this is when Eq. (9) holds for substructures A, B, C, D .

Here A is connected to B both through $I(A, B)$ and through $I(A, D), I(D, C), I(C, B)$. If we remove interlocking $I(A, B)$, A is still connected to B . In general, we can have

$$\Pi_{i=1\dots m} I(H_i, H_{(i+1) \bmod m}) \quad (11)$$

for substructures H_i . Here mod is the modulo operator. It is common to see such circular interlockings with $m = 3$ or 4 in PDB structures. An example of this is shown in Fig. 6(b).

In an assembly the efficiency of the nonpolar SCs in contributing to blocking is further enhanced as a SC can participate in interlockings involving multiple substructures. The DOF reduction through core assembly and the structural strength of a core assembly will be discussed after we have introduced its truss representation and associated properties.

III. TRUSS REPRESENTATION OF CORE ASSEMBLIES

To understand the strength and stability of the core assembly, the equilibrium of the structural members must be

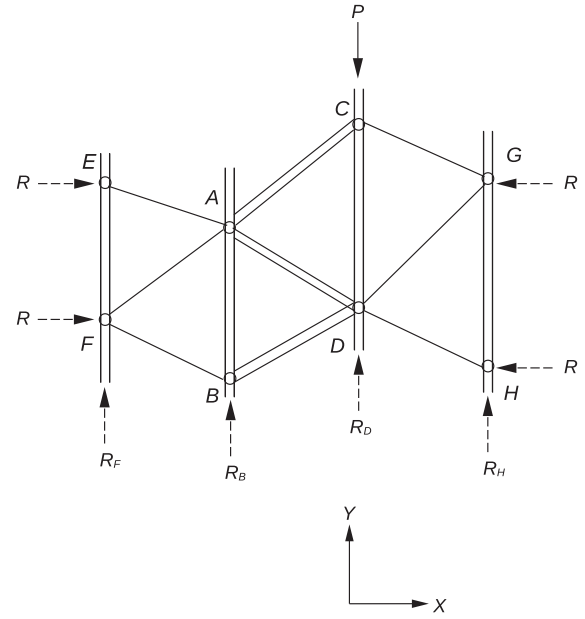


FIG. 7. A truss representing three substructures in interlocking. The notation convention is the same as in Fig. 3.

analyzed. The protein *native structure* is then identified by analyzing the stable equilibria of all potential core assemblies and finding the one that has the highest failure load. This can be done by applying basic principles in mechanics and by borrowing, with modifications, some tools from conventional structural analysis. A particular tool that we use in our static model of protein structure is the *truss representation*. One gets the image of a truss from truss bridges or internal frames of factory shops. An example of two-dimensional (2D) truss is shown in Fig. 7. That of a 3D truss is shown in Fig. 5. In structural analysis, a truss is treated as consisting of bars, i.e., line segments, joined by joints, i.e., points, and supported at the joints [61,62]. In general a bar can only carry axial forces but can carry both compression and tension. If a bar can only carry compression, it is a *compression-only* member. All the internal forces are through the bars and equilibrium can be established at each joint. That is, formally the equilibrium equations are

$$\text{For every joint } j \text{ in truss for every dimension } d \in \{x, y, z\} \\ \sum F_{j,d} = 0. \quad (12)$$

Thus, the analysis can avoid more complicated actions on structural members such as shear forces or bending moments. This makes truss an ideal medium for building a simplified model of structures, including protein structures.

A. Interlocking represented in a truss

When two substructures are interlocked, a load applied on one will have effect on the other. As a result, an interlocking has the additional function of diverting the load aside from fixating the structure. To appreciate this, we can model each blocking between an interacting pair of SCs as a bar in a truss. As such, a residue including both SC and MC units will be a joint in the truss. Later, we may refer to a joint as a SC. This

is because the intensities of bar actions are often calculated on the basis of SC interactions. As shown in Fig. 3, we have four SCs similar to Fig. 2(c). Bars AB and CD are part of MC, while bars AC , AD , and BD represent (mutual) blocking.

In analyzing a plane truss, the *basic equation* [61–63,70] is

$$2j = b + r. \quad (13)$$

Here j is the number of joints, b number of bars, and r the number of reaction forces. This equation can be interpreted in two ways. First in 2D each point has two degrees of freedom, totaling $2j$. Adding each bar or reaction potentially constrains one degree of freedom. For the equilibrium the two should be equal. A second interpretation is to see this as solving a linear equation system with $2j$ equations in which each joint will have two equations for x - and y -dimensional degrees of freedom [consider Eq. (12) in 2D]. If we have exactly $b + r = 2j$ unknowns, we will likely have a unique solution. If the sum $b + r$ is smaller, the truss is not fully constrained and the equation system will not have a solution. If the sum $b + r$ is larger, then the structure is overdetermined or *indeterminate* [61,62,74]. The extra $b + r - 2j$ forces and reactions will be considered as *redundants*. We will have to resolve the indeterminacy through additional geometric compatibility constraints.

Turning to the example on Fig. 3, we apply the equation $2j = b + r$. Assuming all bars can be both tensile and compressive, one can verify that with just three reactions, the truss will be solvable and determinate. With total six reactions, this truss is indeterminate.

If a tandem interlocking can be viewed as part of a plane truss as above, then a cross interlocking (as shown in Fig. 4) can be modeled as a space truss with the blocking pairs treated as bars. The basic equation for space truss is $3j = b + r$ with similar interpretations as for the 2D equation. With 4 joints, $3j = 3 \times 4 = 12$. There are six bars. Six support reactions are needed, which is exactly the number of degrees of freedom for a rigid body in space.

We can analyze the equilibrium of this truss like we have done with the 2D truss (Sec. II B) and see how the load is distributed and how large a force an intermediate bar will carry. We will not go into the details now. It is sufficient to point out that since at each joint there are two bars coming from the other substructure, the burden on each of the bars will be less. Furthermore, the load that is axial in one substructure, e.g., AB in the figure, becomes lateral in the other. This has consequences in load distribution.

Constructing a truss representation for a core assembly

To analyze the equilibrium properties of a core assembly, we can map it into the truss representation. The construction of this truss is by first separately constructing the partial truss for each substructure. In this construction we must make sure each substructure is itself stable, that is, satisfying $3j = b + r$ where $r = 6$. When $j \geq 3$, MC bars that only connect sequential neighbor residues will not be enough (then $b = j - 1$). A large number of extra MC bars need to be added. However, this does not present an operational problem. In many cases as a sufficient condition a simple procedure can be followed: If a space truss can be constructed by successively adding

tetrahedrons to a starting stable truss, then it will be stable. Note that this is exactly adding three bars per one joint, thus satisfying the basic equation at each step. We then can add the compression-only bars which are the blocking forces from neighboring substructures.

Aside from LJ forces, a wide range of forces in interlocking are represented as bars in the truss representation. These include h-bonds, π stacking, disulphide bonds, and salt bridges. The first three force types usually have the counterparts of nonpolar SC-SC compression-only interactions yet themselves are capable of both tension and compression. Because of this, in the truss, we can simply change the compression-only bars to generic bars, i.e., both tensile and compressive. Many salt bridges and some h-bonds are far apart from their associated nonpolar lockings. For them specific bars may be added to the truss and in the equilibrium analysis. Destabilizing forces can be included in the truss too. For example, an unneutralized charged group will be a load on a joint that is created for the charged SC. If there are j joints in the assembly, this will result in $3 \times j$ equations. Geometric and constitutive compatibility equations are often needed to resolve indeterminacy. For this the stiffness calculated through Eq. (3) will be needed.

B. Load distribution problem for protein core assemblies

We are interested in how strongly a core assembly can resist a potentially disruptive external force, i.e., load. This capability can be measured in two ways. First, on the blocking side, how strongly each bar is engaged in. This is described in Eq. (1). Second, how a load on a single substructure can be distributed onto all the substructures in the core assembly. Furthermore, in the process of force transfer, on which bar the force is the largest. Now we address the second measure.

For a conventional truss, this can be solved by solving the equilibrium equations if the truss satisfies the basic equation $3j = b + r$. When $3j < b + r$, the truss is indeterminate. Still, mechanics has a well-established procedure on the basis of energy conservation and geometric constraints to solve the system. A rudimentary understanding of this procedure is essential in following how our model operates. However, because it is not in our model per se, it is presented in the Supplemental Material [66].

Briefly, the procedure is formulated as an optimization problem but is then reduced to solving a linear system of equations. Relative to the equilibrium equations, this is a second set of equations, called *compatibility equations*. They relate the deformations and associated strain energy changes. For a bar in conventional truss, the strain energy arises from the stiffness associated with well-known elastic modulus. For the blocking bar in the truss representation, it is the stiffness defined in Eq. (3).

Assume $n (= b + r)$ is the number of bar forces and reactions, i.e., unknowns in the equilibrium equations of the truss, and $m (= b + r - 3j)$ is the *degree of indeterminacy* (or the number of redundants). As the solution process (shown in the Supplemental Material [66]) will demonstrate, solving the indeterminate truss requires (1) solving $m + 1$ linear systems of equilibrium equations of order $3j = (n - m)$ in the form of $Mx = b$ and (2) solving a linear system of equations of

the form of $M_1 \xi = \beta$, where M_1 is an $m \times m$ symmetric matrix. As a result, computationally the complexity of the equation solving is at least $O([\max(3j, m)]^3)$.⁵

In a conventional truss some indeterminacy will be present. But the indeterminacy is necessarily much higher in a truss representation of a core assembly. This is because of the compression-only nature of its blocking bar. Assume there are n_s residues in the core assembly, there will be the same number of joints in the truss. For each joint, there will be several blocking bars connecting to it. From Sec. II B, because of the compression-only nature of these blocking bars, we know that for each of these bars usually a reaction is needed to balance one, two, or three components of the bar force. Assume on the average the reaction number is $1\frac{1}{2}$ per joint, there will be total $3/2n_s$ reactions. We know a self-standing truss only needs six external forces to fix its position in space. Thus, when we have this many reactions, we will have a much larger r in the basic equation $3j = b + r$. This will be a structure with high degree of indeterminacy, which implies high computational complexity. In searching for the native structure, the search program needs to examine a huge number of core assemblies, each with large number of nonpolar interior contacts. This high order of complexity can be a serious burden.

In the Supplemental Material [66] it is shown that the simple trusses therein exhibit the following patterns of behaviors: (1) a load on one substructure is distributed to nearby substructures; (2) a 3D truss enjoys more even load distribution than a similarly constructed 2D truss; (3) it seems the more densely the support reactions are located near the load, the more locally distributed is the load. With a similar high density of reactions it can be expected that the protein core assembly will behave similarly.

C. Structural strength of core assemblies: Failure load

As the final step in describing our static model of protein structure, we define its structural strength or stability. By that we mean the load at which the structure fails. This can be approached through the equilibrium equation system of the structure, i.e., of the core assembly [Eq. (12)]. In setting up this equation system a *unit* test load is placed in the axial dimension in both directions of each substructure. The partial load carried by each *fixating force* bar in the truss, e.g., blocking, h-bonding, or other force in interlocking, can then be solved. This is checked to find P , a scaling of the unit load, that exceeds the structural strength of the bar, i.e., breaks it. We can choose the load that breaks the “key” interactions in *all* the interlockings, which means breaking the strongest fixating interactions, as the *failure load* for the structure.

For the truss of the core assembly under load P , a solution bar force $F_{j_i, P}$ can be found for every fixating bar indexed by j_i of interlocking i . Let B_i be the set of bar indices for interlocking i . Let $f_{j_i, \max}$ denote the strength of bar j_i as defined in Eq. (2). Then, failure load of the structure is the

minimal load P that can produce a bar force $F_{k_i, P}$ exceeding the bar strength $f_{k_i, \max}$ for the strongest bar of each of the interlockings. Formally,

$$P_{\text{failure}} = \text{Minimum } P \text{ such that [for bar set of every interlocking } B_i \text{ (let } f_{k_i, \max} = \max_{j_i \in B_i} f_{j_i, \max}, F_{k_i, P} \geq f_{k_i, \max} \text{)]}. \quad (14)$$

This definition does not require every bar force F_{j_i} exceeds the strength of the bar j_i , $f_{j_i, \max}$. That requirement is only on the single bar that is strongest in the interlocking. The assumption is that the experimentally determined “key” interaction is also the strongest among those in an interlocking.

Recall that $f_{k_i, \max}$, based on Eq. (2), is an empirical quantity. It is calculated from the current resultant LJ force between two SCs. It is dependent on the distance and orientation between the two. But, it should also be dependent on how the two SCs are supported laterally and how the SCs that support them are themselves supported. In the former case, the possible failure is likely abrupt. In the latter case, the failure would be more gradual, as in a slipping-off scenario, which is like buckling as described in Sec. VI B.

Now that the structural strength (i.e., failure load) for a core assembly is defined, the *native structure* in our static model will be the conformation whose core assembly has the highest structural strength in comparison with all other conformations.

IV. COMPARING INTERLOCKING FEATURES OF CORE ASSEMBLIES

In many practical applications, we are comparing the stabilities of two conformations, typically deciding if one is the native structure and the other a decoy. In that case, we only need to know the *relative strength* of the two. This approach is supported by the nature of protein structure. Statistical mechanics of protein folding dictates that a native structure must be *substantially superior* to all (or most) other conformations. This implies that there can be salient features or measures that distinguish the native structure at a qualitative level. This further implies we may apply our truss representation to calculate the patterns and intensity of substructure interlocking without solving the system of equilibrium equations. This can be more desirable than just finding a failure load. First, by identifying patterns of interlocking with higher structural strength and stability, the search for native structure can be more efficient [75,76]. Second and more important for testing our model, it can avoid adoption of potentially controversial or uncertain parameter values.

A. Redundancy in core assembly: Duplicate and circular interlocking

Redundancy is generally considered a factor that strengthens structures [61,62,71,74]. The benefit of redundancy is first and foremost in sustaining the equilibrium. That is, if a particular structural member fails, a redundant member that provides support at the same junction can come in and the structure can still stand. In our model, it also has another quantifiable benefit: the proportional load reduction when adding an additional interlocking.

⁵Despite the multiple of $m + 1$, because only the b term in the linear equation systems $Mx = b$ changes, matrix M only need to be factorized once, e.g., through LU factorization. In fact, since M is a sparse matrix, the complexity may be as low as $O[(3j)^{3/2}]$.

In a core assembly, redundancy can take many forms. We mention two: duplicate and circular interlocking. Circular interlocking is introduced earlier [Eq. (11) and Fig. 6(b)]. By duplicate interlocking, we mean that the same pair of substructures are interlocked through two sets of interlocking actions. They can be of different types, e.g., one nonpolar interlocking and another salt-bridge interlocking. But, they can also be just two nonpolar interlockings. We may formulate for the former

$$I(H, H') \text{ and } I_{\text{salt bridge}}(H, H'),$$

where the shorthand is used and the subscript for the second I indicates the type. For the latter, we may write

$$I(H, H') \text{ and } I'(H, H') \text{ and } I \neq I'.$$

To accurately know how a particular duplicate setup can change the force distribution, the system of equilibrium equations must be solved. But, we can appreciate the scale of change by examining some simple scenarios. For a duplicate interlocking, we arbitrarily assume that one interlocking I_j , with stiffness K for the blockings [as defined Eq. (3)], is “original” and that the other I_{j_1} , with stiffness k for the blockings, is redundant and added later. With a simplifying assumption that the two interlockings proportionally share the load, it can be shown that the load on I_j becomes

$$P'_j = \frac{K}{K+k} P_j.$$

If, alternatively, I_{j_1} is assumed to proportionally share the load with all the interlockings, assuming the total load is P , the load on I_j becomes

$$P'_j = \frac{P}{P+P_{j_1}} P_j. \quad (15)$$

This latter equation also applies to redundant interlocking arisen from circularly interlocked substructures.

B. Concentrated interlocking assembly

As mentioned in Sec. II D, with high coordination substructures we can have concentrated interlocking assembly. Thus, when two core assemblies are otherwise equal, we can compare their mean coordination to see if one is superior. If the two mean values are not significantly different, we can check if one has a much larger maximal coordination number since that *usually* would enable the core assembly to do better in efficiency, steadiness, and strength. But, there can be complicating cases which we show below.

We can have two assembly patterns related to the familiar 4-helix bundle. The four helices can have one in the center (H_1) and the other three interlocking onto it. Then additionally two out of the three (H_2, H_3) form redundant interlocking as shown in Fig. 6(a). Alternatively, there can be a circular interlocking with each forming two interlockings as shown in Fig. 6(b). The former case can be expressed as

$$\Pi_{i=2...4} I(H_1, H_i) \text{ and } I(H_2, H_3).$$

The assembly will have a mean coordination number of $\langle C \rangle = 2$. If we arrange the coordination numbers in decreasing

order, it will be $\langle 3, 2, 2, 1 \rangle$. The latter case will be

$$\Pi_{i=1...4} I(H_i, H_{(i+1) \bmod 4})$$

the assembly will have similarly $\langle C \rangle = 2$. But, the sequence of coordination numbers is $\langle 2, 2, 2, 2 \rangle$. By adding a new interlocking, the two cases will become identical. This is shown as dotted line in each figure. Formally, we have

$$\Pi_{i=1...4} I(H_i, H_{(i+1) \bmod 4}) \text{ and } I(H_1, H_3). \quad (16)$$

Mean coordination number is $\langle C \rangle = \frac{5}{2}$. The sequence is $\langle 3, 3, 2, 2 \rangle$. Many 4-helix bundles, e.g., 2MHR, have this configuration. 1AEP with five helices has a similar configuration. Clearly, before adding the fifth interlocking, the two assemblies have nearly equal measures in terms of coordination number. The one that is obtained through adding an interlocking [Eq. (16)] is superior to both original ones.

C. A longer helix vs two short helices from the same chain segment

We are concerned with two core assemblies where one has a helix H° but the other breaks the same chain segment H° into two helices with an intermediate loop, $H^\circ = H * L * H'$. Here each symbol denotes an amino acid chain segment and the asterisk denotes concatenation. As a consequence, if the helices are to get the same near minimal number of interlockings of 2 (see Sec. VI C), we would have

$$I(H, H_{j_1}) \text{ and } I(H, H_{j_2}) \text{ and } I(H', H_{k_1}) \text{ and } I(H', H_{k_2}),$$

where j_1, j_2, k_1, k_2 are indices of substructures other than H and H' . In contrast, for H° , we may have

$$I(H^\circ, H_{i_1}) \text{ and } I(H^\circ, H_{i_2}),$$

where i_1, i_2 are similarly indices of substructures.

Breaking up the chain segment H° may be necessary when the amino acid distribution pattern changes along the chain. When a helix straddles both interior and exterior, the ideal pattern of nonpolar amino acid distribution is typically $r, r+3, r+4, r+7$. This way they are on the same half of the cylindrical surface of the helix, which we will refer as *face*. If two consecutive such patterns can line up on the same face of a helix, then it may be advantageous to not break up the chain segment. On the other hand, if the two patterns will end up on opposite face of a helix, then it would be more advantageous to break them into two.

However, starting a new helix unnecessarily can be a waste. When H, L, H' can place their polar or nonpolar residues separately on the same faces and when the length of H plus H' is not too long for the total chain length (so that the core assembly can fit a relatively globular shape), H° in the assembly is more stable than H, H' being in an otherwise identical assembly. This is because (1) an additional helix will add six extra degrees of freedom, which must be handled by interlocking. (2) The additional helix will compete for SCs to interlock with a scarce resource. (3) If the additional helix, e.g., H' , does not align with H , i.e., together they take up the same continuous space, then H' will cause a less regular packing in the assembly.

There can be the reverse of the phenomenon that we just discussed. That is, for the same $H^\circ = H * L * H'$, breaking up

H° is necessary: Otherwise, the nonpolar faces and polar faces of helices may be mixed up and an unneutralized charged group may be buried.

When an unneutralized charged group is on a substructure of the core assembly and is buried in the deep interior, in terms of the equilibrium, we can view it as a *polar load* P_p , being applied on the substructure. In the Supplemental Material [66] we explain that its presence will most probably reduce the interlocking intensity.

D. Implementations

The program for extracting and comparing protein core assemblies is implemented in SBCL [77], a version of common Lisp, and can also run in Maxima [78]. Most of the results reported below can be directly calculated from the formal definitions. Many of the critical parameters are mentioned in the Results section. Here we just explain two implementation issues: how we calculate interlocking intensity and how we determine if a SC is interior versus if it is “exposed” to solvent.

The evaluation of blocking intensity, f_b in Eq. (1), between two SCs is mainly based on geometric measures such as distances and angles rather than a force field. A pair of nonpolar atoms, including nonpolar hydrogen atoms, is assumed to be in repulsion when the gap distance between them, i.e., the difference between their center-to-center distance and the sum of their radii, is under a threshold d_{thresh} . Two SCs are considered in interaction when their respective atoms are in interaction. When the two interacting SCs are from two neighboring substructures, if the angle between the vector of this interaction and the axis of one of the substructures is smaller than a threshold, a blocking exists. Note that since there are one-to-one mappings between gap distance and center-to-center distance and between the angle and the axial component of the force in Eq. (1), theoretically we can have a mapping from the threshold δ in Eq. (1) to d_{thresh} and an angle threshold. Yet, we chose to implement the calculation using this distance and angle measure. This is to avoid biases when adhering to a particular FF, which in turn is because the C_{12} and C_6 (alternatively σ and ϵ) values for LJ potential are substantially different among CHARMM, Amber, Gromos, and OPLS [46–50].

The choice of d_{thresh} is critical to the validity of the calculation results on interlockings and core assemblies. A sign of a poor parameter choice (or even the failure of the model) would be that the results change substantially with an insignificant change of the parameter value. We have determined that 4.15 to 4.35 Å is a distance range appropriate for d_{thresh} and that the angle threshold can be neglected in the context of core assembly calculation. Both are discussed in details in the Supplemental Material [66] (Sec. VII B).

Furthermore, because (1) the orientation between the interacting pair SCs r and r' fluctuates and (2) the presence or absence of a blocking or locking is a discrete measure, in ranking intensities of interactions, we quantify interactions at a few discrete levels rather than giving them an apparently precise continuous number. Once two nonpolar side chains are considered in a compression on the basis of closest atomic distance, a force level is assigned. It ranges from 1 for a distance in the interval of 4.25 to 4.35 Å to 8 when the distance

is less than 3.4 Å. The level is statistically determined from average LJ forces at close atomic distances.

The intensity of the interlocking between two substructures is the sum total of all blockings involved. Namely, we first determine if there are enough number of blockings to qualify two substructures as in interlocking. Once this is affirmed, then all the blockings are counted. The intensity is included in two numbers: the total number of blockings on a substructure and the sum of all the levels of all the blocking forces.

To determine if a charged or strong polar group destabilizes a core assembly, we must first determine if it is in the interior of the conformation. If it is, we next need to determine if there is not a large enough opening to the exterior near the group to neutralize it. Lee and Richards algorithm [69] calculates solvent accessible area. But, this is not necessarily whether a group is exposed to solvent. Some protein structures have a large interior void. A standalone charged or strong polar group that is left in such a void will be considered solvent accessible but still energetically strongly unfavorable. To locate such cases, we must first be able to identify those that are buried in the interior. Sometimes the interior and exterior have a clear boundary, e.g., in the case of beta sheet involved structures such as beta barrels, beta sandwiches, and alpha-beta-alpha structures. In other cases, we directly test if such a group is fully surrounded by protein atoms. For the 4π solid angle around the centroid of the group, we evenly orient 16 vectors in space and check the cone of solid angle of $4\pi/16 = \pi/4$ that has the vector as the cone axis. If in every orientation this cone will contain some protein atoms, the group is in the interior. Next, we find all the rectangular openings on a spherical surface centered at the group centroid in which the opening is delimited by atoms in the protein. If the opening is large enough, then the group is considered *exposed* to the solvent and possibly neutralizable. In a survey of a large set of native structures, all buried charged or strong polar groups pass this test. Note that the second procedure subsumes the first but it is much more computationally intensive. Thus, the first one is necessary for efficiency.

V. RESULTS IN COMPARING CORE ASSEMBLIES

In this section, we show that interlocking-based stabilization mechanisms are widely observed in protein structures. Next we show that one can use constraints on the general properties of core assemblies to screen decoys. Depending on specific sequences, this can sometimes successfully prune most or even all decoys. However, if we are intent on practically fully detecting and distinguishing decoys, other conformational properties must also be considered. In the Supplemental Material [66] (Sec. VII C), we show that in the case that a decoy is not pruned by checks on assembly properties and the decoy conformation is of satisfactory quality [79], e.g., free of severe steric conflicts, it can be pruned by local energetical concerns like excessive prolines in a helix and global ones like burying unneutralized charged or strong polar groups in the nonpolar interior.

A. Various interlocking types and assembly patterns in native structures

We first present data on how interlocking-based mechanisms cover the various types of structures, e.g., helical,

TABLE I. Distribution of various types of interlockings in helical structures. The columns of the table are as follows: L indicates chain length, S the number of substructures, i.e., helices and h-bonded beta strands, N_S the residues in the S substructures. N_S/L (%) then is the percentage of residues in substructures relative to the total chain length. These are all PDB structure data. s is the number of substructures that are interlocked by various mechanisms. s/S (%) is the coverage percentage. n_s is the number of residues in the covered s substructures. n_s/N_S (%) is the percentage. The next five columns list the interlockings by type. n_{il} is for nonpolar interlocking. n_{es} for interlockings exclusively by salt bridges and strong h-bonds. n_{hpin} by hairpins and only by hairpins. n_{ssb} by ssbonds. n_Σ is the sum total of the preceding four columns. f_Σ the total number of LJ and ES forces in the interlocking. π -stacking-effected interlocking is not separately counted. This is because when two aromatic rings are in close distance and in proper orientation, they will have both LJ and π -stacking interactions and often already have nonpolar interlocking. Note that $n_\Sigma \geq s - 1$.

	L	S	N_S	N_S/L (%)	s	s/S (%)	n_s	n_s/N_S (%)	n_{il}	n_{es}	n_{hpin}	n_{ssb}	n_Σ	f_Σ
2CRO	65	5	43	66	5	100	43	100	2	2	0	0	4	11
1UTG	70	4	55	79	4	100	55	100	4	0	0	0	4	14
3UCY	74	6	57	77	5	83	52	91	5	1	0	0	6	28
5CPV	108	9	76	70	4	44	42	55	4	0	0	0	4	23
2MHR	118	4	85	72	4	100	85	100	3	2	0	0	5	30
1TAM	120	7	82	68	4	57	68	83	3	1	0	0	4	16
2CCY	127	4	101	80	4	100	101	100	4	0	0	0	4	25
1AKI	129	12	79	61	7	58	57	72	3	0	0	4	7	18
1ECA	136	8	114	84	8	100	114	100	8	2	0	0	10	46
1MBC	153	9	127	83	9	100	127	100	7	4	0	0	11	55
1AEP	153	5	134	88	5	100	134	100	5	2	0	0	7	53
153L	185	11	119	64	7	64	103	87	5	2	0	2	9	38

beta-sheet, and alpha-beta structures. The individual structures are chosen for their diverse lengths and packing characteristics but also for how frequently they appear in literature. This should provide a sense of how prevalent each interlocking pattern exists in protein structures and how large a portion of the protein structures these patterns can already cover. Next, we show the comparison between the native structures and decoys. This is based on feature comparison that we have introduced in the previous section. They will be shown to be able to tell whether one core assembly is generally more stable than the other.

Tables I and II show measures and characteristic features of interlocking in 12 individual helical structures. The top half of Table III shows those of 50 various types of native structures. The features of individual beta or alpha-beta structures are similar and the corresponding tables are included in the Supplemental Material [66]. A few structures originally contain ligands. Some ligands are metal ions. Some are larger, such as hemes. The conformations without the ligands are included when it is determined that the PDB structures are not far from their apo form, through comparison with NMR structures of the same or similar sequences or through other means. Some homologous structures are intentionally added to show that their core assemblies have strong resemblance in most but not all cases. We also note that for simplification we have used the same force level threshold for all the protein sequences, which may miss some interlockings for sequences with compositions that generate low intensity blockings.

It can be seen from the tables that the core assembly for most of proteins can include a large percentage of substructures, even using the same force level threshold. For helical structures, the overwhelming majority of interlocking is non-polar. For beta sheet structures, salt bridges and h-bonds also take a large percentage. This is obviously due to the h-bonding between strands. The role of hairpin is often to reinforce

already established interlocking from nonpolar contacts or ES interactions. Thus, in the column of n_{hpin} there are many zeros.

TABLE II. Four features of interlocking in helical structures. “Duplicate interlocking” is the count of substructure pairs that are interlocked by more than one mechanism, e.g., by both a nonpolar interlocking and a few salt bridges. In a conformation, some substructures pair with more substructures than others. They serve as hubs of packing, making the interlocking more concentrated and stronger. “High coordination substructures” list the number of pairs that these substructures make. Only two higher coordination levels are included. When we see only one number, it is because there is no substructure that pairs at the next level. That is, the packing is highly concentrated. “Cross interlocking” lists the number of cross interlockings. “Nonduplicate redundant interlocking” is the difference between the number of interlocked substructure pairs n_Σ (refer to Table I) and the minimum number of such pairs for interlocking all the covered substructures, $s - 1$. Normally it is the number of circular interlockings (Sec. IV A).

	Duplicate interlocking (n_d)	High coordination substructures	Cross interlocking	Nonduplicate redundant interlocking
2CRO	0	(3 2)	0	0
1UTG	2	(3 2 2)	0	1
3UCY	1	(3 3 3 2)	1	2
5CPV	3	(2 2)	0	1
2MHR	3	(3 3 2 2)	0	2
1TAM	1	(3)	0	1
2CCY	4	(3 2 2)	0	1
1AKI	5	(4)	0	2
1ECA	3	(4 4 4)	1	3
1MBC	4	(5 4)	2	3
1AEP	2	(3 3 3 3 2)	1	3
153L	4	(5)	1	3

TABLE III. Averages of measures of interlockings for various sets of native structures or decoys. The upper half shows native structures, the lower half decoys. A PDB structure identifier with “-DC” indicates decoys of that PDB structure. We use $\{\sigma\}$ to designate the set of structures of concern for each row. The first column, $|\{\sigma\}|$, is the number of structures. The next six columns are the same measures as in Table I and in corresponding table for beta structures in the Supplemental Material [66], only that averages are taken over the set. A new addition of measure is n'_{hpin} . This is for all the interlockings that contain hairpin interlockings, including those that duplicate interlockings for the same substructure pair through different mechanisms, particularly nonpolar interlocking. n_X is the number of structures in the set where there is a cross interlocking. $n_X/|\{\sigma\}|$ is the ratio of structures that contain cross interlocking. C_{max} is the average of the highest number of interlocking pairs a substructure can make in a structure in the set. n_{dp} is the average of the number of duplicate interlockings.

	$ \{\sigma\} $	N_S/L (%)	n_s/N_S (%)	s/S (%)	n_{il}/n_Σ (%)	n_{es}/n_Σ (%)	n_{hpin}/n_Σ (%)	$n'_{\text{hpin}}/n_\Sigma$ (%)	$n_X/ \{\sigma\} $ (%)	C_{max}	n_{dp}
HELIX	33	73	87	92	75	20	1	5	55	3	2
β	12	70	78	84	56	38	2	7	15	3	2
α - β	5	67	74	78	52	34	3	3	20	3	2
ALL	50	71	82	87	68	26	1	5	40	3	2
MUSTER	22	80	77	80	70	25	5	9	36	2	1
1CTF-DC	19	61	46	57	45	55	0	0	0	1	0
2CRO-DC	23	66	49	46	88	12	0	2	9	1	0
1EH2-DC	26	63	73	83	89	8	3	5	58	3	1
2FQ3-DC	72	71	75	75	100	0	0	0	100	2	0
3LDC-DC	86	81	82	82	98	0	2	9	10	2	0
4HKG-DC	34	75	57	69	77	12	12	12	20	2	1

Table III will show that as reinforcements, it contributes a considerable percentage. A substructure interlocking usually involves at least three forces. But, the same substructure pair may be interlocked by multiple mechanisms as indicated by “duplicate interlocking” column. Furthermore, there can be many more nonpolar compression pairs or ES interaction pairs between two substructures. Thus, almost in all cases $f_\Sigma > 3n_\Sigma$ (refer to Table I and the corresponding table for beta structures). These are local support redundancies. To fixate a substructure, with the constraint of being confined in tightly packed globular environment, the most effective is to restrain it in its most likely displaceable axial dimension, for which one interlocking is enough. Thus, minimally $n_\Sigma = s - 1$. But, more interlockings exist. As Sec. VIC will explain, these are for fixating all the substructures in the three-dimensional space. Here we can view them as redundant relative to the minimal amount. This is indicated in the column “redundant interlocking.” If all substructures are packed in a single core, this number should be $n_\Sigma + 1 - s$. In the rare case that there are two cores, it would be $n_\Sigma + 2 - s$. The helical structures are arranged less regularly than beta structures. Thus, their “high coordination substructures” can have higher contacts. For a similar reason, they are more likely to have cross interlockings.

B. Comparing assembly features between native structures and decoys

Table III gives an overview for large sets of native structures and decoys. The native structures are chosen for mostly nonhomolog ones. They are divided into three groups. The coverage (s/S) is very high for every group. There is a contrast between helical and beta structures. Helical structures rely more on nonpolar interlocking than beta-sheet ones. Beta-sheet structures rely more on ES interactions, i.e., salt bridges and h-bonds. $n'_{\text{hpin}}/n_\Sigma$ is the ratio of hairpin interlocking relative to all interlocking. Clearly, beta-sheet structures use more

of it. On the other hand, helical structures have many more cross interlockings. In fact, more than half of them contain one. In terms of concentrated interlockings on central substructures, they are all the same with an average value of three interlocking pairs. They are also identical in having multiple interlocking mechanisms on the same interlocking pairs, with an average value of 2.

The decoys are in the lower half of the table. We have included six sets of decoys corresponding to individual PDB structures. 1CTF, 2CRO, and 1EH2 are decoys from the Decoys ‘R’ Us database [42,43]. 2FQ3, 3LDC, and 4HKG from the MUSTER decoy database [45]. The former is generated by conformational search. The latter is through threading. We have also included a separate “MUSTER” set that includes decoys corresponding to 11 PDB structures from the same “MUSTER” source. These are decoys with good overall measures from the full set of all decoys that we describe in Tables IV and V.

The individual decoy conformations are first selected based on dissimilarity in RMS deviations (RMSD), both from the native and from each other. Here not all RMSD-wise close conformations are mutually excluded. For example, some decoy conformations have excessively high steric conflicts. A separate conformation will be generated by minimizing the original conformation [46].

The column of N_S/L is a feature independent of the interlocking properties. As shown, the decoys have as many substructures as the native structures. This should be expected given how they are generated. In fact, many decoys are constructed to have nearly identical secondary structure assignments as the native structures. The coverage, that is, the ratio of substructures interlocked in the core assembly versus the total set of substructures, both in terms of number of substructures (s/S) and number of residues (n_s/N_S) are generally lower than the native structure. But, the numbers are comparable. This in a sense is good news. It implies that forming interlockings in packing the substructures is not too

TABLE IV. Summary of the comparison of native structures (N) and decoys (D) on the basis of feature comparison. *Original decoy set* is the whole set of decoys for a PDB structure, taken from a decoy database. Among them only those that are RMSD of 7.0 Å or larger from the native structure are *tested*. Note that this set is a superset of the set for the same sequence in Table III. There are two differences: (1) When taking averages there, decoys that do not have a core assembly are excluded; (2) there the tested decoys are also filtered by pairwise RMSDs. Only those with pairwise-RMSD of 7.0 Å or larger are included. *Winners* indicates the number of conformations that are superior in the comparison. Some decoys simply do not have a core assembly. Some are so close to the native structure in the features that it has to be called a *tie*. They are listed in the last two columns.

	<i>L</i>	Original decoy set size	Tested decoy set size	Winners	%	Decoys with no assembly	Tie
1CTF(N)	68			26	100	7	
(D)		500	26	0	0		
2CRO(N)	65			26	100	3	
(D)		500	26	0	0		
1EH2(N)	95			54	69	6	6
(D)		11400	78	18	23		
1APC(N)	106			112	92	28	5
(D)		300	122	5	4		
1CY5(N)	92			121	100	14	
(D)		300	121	0	0		
1FK5(N)	93			120	94	54	1
(D)		300	127	6	5		
1LWB(N)	122			130	99	15	1
(D)		300	131	0	0		
2FQ3(N)	85			107	79	29	8
(D)		300	135	20	15		
2J9W(N)	101			127	100	3	
(D)		300	127	0	0		
3FYM(N)	82			128	100	73	
(D)		300	128	0	0		
3LDC(N)	82			113	92	13	9
(D)		300	123	1	1		
4A56(N)	93			123	100	64	
(D)		300	123	0	0		
4GMQ(N)	92			130	100	34	
(D)		300	130	0	0		
4HKG(N)	80			124	99	27	1
(D)		300	125	0	0		
4J1P(N)	114			126	99	9	1
(D)		300	127	0	0		
4LUP(N)	91			122	100	14	
(D)		300	122	0	0		

subtle a task that current conformation generation methods can not perform. The lower level of interlocking on the decoy side may have more to do with the choices of packing patterns than the difficulties of locally forming strong interlocking geometrically.

The decoys, except 1CTF decoys, are mostly helical. Thus, they behave very much like the helical group of native structures in the top row. Their interlockings are mainly nonpolar interlockings. Only 1CTF decoys have a larger percentage of ES interlockings. Cross interlocking is more frequent in

helical structures, but it is sequence specific. In this lineup, we see that 2FQ3 decoys all have cross interlockings but 2CRO decoys have only 9% that do. The MUSTER decoy set has a number that is in-between, 36%. This is significantly worse than the native helical structures but close to the average for all native structures. The average number of high coordination interlocking substructures and duplicate interlocking counts are both inferior to those of the native structures.

We have applied our model and feature comparisons for ranking structural stability to the pruning of decoys. Two conformations are first compared in how their interlockings can cover the substructures to the same extent. If not, assuming the interlocking strength of the two are similar, e.g., similar number of interlockings and similar number of interlocking forces, the one that covers more substructures prevails. If the two cover similar number of substructures or residues, usually within 10%, they will be compared in measures like number of interlockings, numbers of high coordination substructures. In a previous section we have shown that when everything else being nearly equal, these can determine if one structure is stronger than another. When all these are nearly equal, as a heuristic we compare the number of interlocking forces. If one is higher than the other by $\frac{1}{3}$ of the total force count, its structure is presumed to be stronger. Another useful measure is to see if a conformation has employed more substructures for the same number of residues in building the core assembly. The justification is given earlier in Sec. IV C.

The comparison results are shown in Tables IV and V. The first table gives a summary. The second details the comparison by individual features. In the listed structures, the first three are from the Decoys ‘R’ Us database and the rest are from the MUSTER database.⁶

1CTF and 2CRO decoys from the first three behave similarly. Because these conformations are generated from scratch, the packing density tends to be lacking relative to the MUSTER decoys which derive their conformations from the templates of actual PDB structures. As shown, quite a few of the 1CTF and 2CRO decoys simply do not form a core assembly. When they do, a large percentage have lower coverage than the native. The same decoys also tend to have fewer interlockings, fewer high coordination substructures, and fewer interlocking forces than the native structure. 1EH2 behaves differently and similar to 2FQ3 from the MUSTER set. They will be discussed together.

The MUSTER decoys are for 13 PDB structures. As shown in the summary table, 2FQ3 and, to a much lesser extent, 1FK5 and 1APC are those that can have a few decoys that have better values in the measures than the native structures. They all win essentially by higher packing density. As a result, they all have larger number of interlockings. Two out of the three have higher number of high coordination substructures as well. For the rest of the decoys, when the sequence composition is not strongly unfavorable for tight packing, the native structures prevail. They win mainly by the same measures, i.e., larger number of interlockings and higher number of high coordination substructures. Note that decoys of most of the

⁶In the MUSTER data base, the 1APC entry here is listed under PDB identifier 256B which is of the identical sequence but is ligated with a heme and other groups. 1APC is the apo form.

TABLE V. Individual feature comparisons between native structures and decoys. The column *Winners* is identical to that in Table IV. It is here to serve as the basis for reference in terms of number of conformations. The rest of columns all have the same format: one column indicating the feature by which the native structure or decoys win, then the next column showing the percentage relative to total winners. “By coverage” indicates that the losing side covers considerably less either in the number of substructures or in the number of residues. The rest of the column labels are self-explanatory. The last two measures are not sufficient to decide whether a conformation is inferior to another. But they are good heuristics for having a guess as to the quality of the conformations. *By not splitting substructures (unnecessarily)* refers to the phenomenon discussed in Sec. IV C, where the losing side breaks up a substructure with coherent faces into two. The last measure indicates that the losing side has significantly fewer LJ or ES forces for the interlockings.

	Winners	By coverage	%	By more substructures	%	By more interlockings	%	By high coordination	%	By not splitting substructures	%	By more force count	%
1CTF(N)	26	19	73	18	69	19	73	19	73			19	73
(D)	0												
2CRO(N)	26	23	88	23	88	23	88	23	88			21	81
(D)	0												
1EH2(N)	54	28	52			47	87	47	87	1	2	40	74
(D)	18	13	72	13	72	13	72	1	6			6	33
1APC(N)	112	77	69	84	75	78	70	78	70			75	67
(D)	5					5	100	3	60				
1CY5(N)	121	101	83			107	88	107	88			107	88
(D)	0												
1FK5(N)	120	63	52	66	55	66	55	66	55			53	44
(D)	6	1	17			3	50					6	100
1LWB(N)	130	111	85			115	88	115	88			115	88
(D)	0												
2FQ3(N)	107	62	58			78	73	78	73			76	71
(D)	20	10	50	3	15	17	85	3	15			7	35
2J9W(N)	127	102	80	111	87	124	98	124	98			124	98
(D)	0												
3FYM(N)	128	55	43	55	43	55	43	55	43			55	43
(D)	0												
3LDC(N)	113	46	41	55	49	94	83	93	82			97	86
(D)	1	1	100										
4A56(N)	123	59	48	59	48	59	48	59	48			59	48
(D)	0												
4GMQ(N)	130	95	73	96	74	96	74	96	74			96	74
(D)	0												
4HKG(N)	124	95	77	96	77	97	78	96	77			97	78
(D)	0												
4J1P(N)	126	113	90			117	93	117	93			117	93
(D)	0												
4LUP(N)	122	92	75	97	80	108	89	108	89			108	89
(D)	0												

sequences behave like 3LDC and 4HKG shown in Table III. The decoys are tightly packed. But, the interlocking number is still short of that of the native structure.

1EH2 and 2FQ3 decoys represent a type of decoys that can have apparent tighter nonpolar packings than the native structures. In general, they achieve this by burying polar or charged groups in the core assembly. In contrast, to avoid burying polar or charged groups in the nonpolar interior, the native structure has made some helices shorter so that some charged SCs are on the loops. This helps the overall stability of the assembly. However, because the feature comparison here does not include those considerations (specifically, polar load mentioned in Sec. IV C), both decoys have quite a few winners and a few that tie with the native structure. As Table V shows, the coverage appears better and there appear to have more interlockings than other decoys.

Specifically for 1EH2, its sequence composition favors beta strands for some chain segments. As a result, 1EH2 decoys are strong in multiple interlocking categories including nonpolar interlocking, ES, hairpin, etc. With tighter packing, this enables it to have better coverage overall than the native and larger interlocking count. On the other hand, as Table III shows, all the interlockings on 2FQ3 decoys come from nonpolar interlocking. The large amount of nonpolar interlocking allows it to have cross interlocking in every decoy. It beats the native structure by the number of interlockings, by coverage, and by the number of high coordination substructures. Naturally, with the sequence composition, 2FQ3 decoys also may have poor core assemblies. In fact, 29 out of 135 decoys do not have one.

For both 1EH2 and 2FQ3, those that avoid pruning by core assembly features can be pruned on the basis of energetical

considerations. The detail is in the Supplemental Material [66].

VI. DISCUSSIONS

The static model proposed here is a departure from consensus thinking about protein stability, particularly relative to the theoretical models that have been developed and widely accepted over the years ([24,80–87]). In the Introduction, we have discussed the need for an alternative model. Here we will discuss the validity of our model in depth. We will first explain that it is consistent with the previous work. In the next few subsections, we will show that the dominant form of interactions in protein interior, the compressions between SCs, has a condition on the stability of its equilibrium, which links our model to thermodynamic hypothesis. We will also show that in the truss representation, the compression-based interlockings can put the core assembly in a stable and determinate state. Limitation of the model and future work are also discussed.

First, while our static model concentrates on analyzing time and ensemble-averaged aspects of protein structure properties, this is not in contradiction to investigating these properties via dynamics and thermodynamics approaches. On the contrary, our model is based on and incorporates the results from the latter. In particular, it is generally agreed that there are two types of determinant factors in protein folding: entropic and enthalpic. But the static model, focusing on the force equilibrium and stability, seems to have ignored the entropic factor. We point out that the assumption regarding the inward normal force and the globular shape, both of which arise from hydrophobic collapse, has captured the effect of the entropic factor. We have to “convert” some physical properties into time and ensemble averages as this is the way to fit them into the uniform treatment of force and moment equilibria. More importantly, by deriving an equation on the buckling condition of blocking actions, we have proven that protein stability in our model is associated with a form of mechanical energy minimum. Thus, it is consistent with the thermodynamic hypothesis.

Second, it has been settled that the protein’s relative stable structure is not due to an intricate gadgetrylike mechanism that “traps” or “snaps” the protein chain into a shape. Our “locking” idea may be perceived as like that but it is not. It is purely based on removal of DOF due to (mainly) compressions. Ours is also a mechanism. But it is so in mechanics sense. It is more like that of an arch or dome. In fact, this makes our model easily reconcilable with many of the observed protein behaviors in folding and unfolding. For example, protein is not well packed in terms of packing density but seems structurally well supported [41]. According to our model this would be like a domed and buttressed church, which does not need to be filled to be strong against its heavy load (mostly its own weight).

Limitations of this static model. Fundamentally, there is no limitation to applying statics to the analysis of the protein structure. This is in the sense that proteins must follow the laws governing all structures: its structural members must be in equilibrium of both forces and moments and the equilibrium must be stable. Because of this, the static model is capable of explaining the geometric features and nuances in

protein structures. For example, evidently the need for even load distribution provides a direct explanation for the widely observed symmetry therein.

As the model stands now, we see several potential shortcomings related to the assumption of linear elasticity of the blocking action and the solid body. A basis of this assumption is surveys on approximating LJ forces for nonpolar SCs in the interior. In one case, through function approximation [88], in a range of [3.0, 3.6] in Å, the LJ value can be approximated by a linear function for about $\frac{3}{4}$ of the points to within 20% of the true value. Many of the structural features do not need a strict linear response. Thus, the error may be acceptable. But this is still to be checked.

Future work. This research is ongoing. The focus of the future work will be the relationship between sequence composition and geometric-mechanical properties of interlocking configurations. Our experience has shown that such relationship can be consequential in rapidly determining the maximum strength and stability of the conformations of a protein sequence. Thus, the search for the most stable conformation can be enormously accelerated [75,76,89]. The program for the equilibrium equation-based calculation of protein structural stability is also currently under development. The aim is to compare the rank orders of the protein stability calculated here with those experimentally obtained.

A. A distinct characteristic of protein structural stability: Compressive support

1. Pure and strong compressive force in the nonpolar interior

A quick check of the interactions between nonpolar SCs in the protein interior will show that these are strong repulsions (compressions in statics terms). And these forces take a very simple form: pure repulsion or compression. In the nonpolar interior this has a structural consequence: it makes the interactions much more uniform and more capable of providing building blocks for a structure. A bar chart is available in the Supplemental Material [66] which shows that the average of nonpolar LJ force is nearly three times of h-bonding and $1\frac{1}{2}$ times of salt bridges.

2. A compression is often steadied by other compressions from lateral directions

It is well known that tension is a *stable* or *steady* action whereas compression is unstable or unsteady [71].⁷ One can call up the images of a ball suspended from a string versus a ball sitting on top of a mound to easily verify that. The arch structure, the prototypical example of pure compression support, despite its equilibrium under an elementary textbook analysis, in practice needs ways to prevent it from sliding off laterally and collapsing. Mortar is commonly used for increasing the friction between stones in an arch. Coulomb

⁷The term “stable” is widely used in structural analysis literature to refer to a property of the equilibrium: an equilibrium can be stable, unstable, or neutral. But it is also widely used to refer to the property of a structure that it is in a static equilibrium state per se. Whenever a confusion may arise, we will use the term “steady” for the former property.

in 1776 derived the specific equations for bracing a masonry arch so that it stays in a stable equilibrium [90].

In the case of protein structure, there is something similar to the mortar for masonry arches. Heavy aliphatic SCs are branched. When engaged in compression, the branches may stagger and thus increase the steadiness of the engagement. In some force fields, e.g., those of Gromos [49], aromatic ring hydrogen atoms have explicit LJ forces. This implies that there will be “studs” on the ring to reduce slippage too. These deterrents to slipping off can be viewed broadly as some minor static “friction.”

More generally and at a larger scale, the steadiness of the compressive interaction can be enhanced by actions from a lateral direction. To prevent columns from earlier mentioned buckling, a common device is a brace added laterally. When two columns are otherwise supported identically, the one braced in the middle [Fig. 1(c)] will have four times of buckling load as the other. This is because Euler’s buckling load is

$$P = \frac{\pi^2 EI}{L^2}, \quad (17)$$

where E is the elastic modulus, I the moment of inertia, and L the length [61–65,71]. That is, it is inversely proportional to the square of the length of the column. A brace in the middle simply halves the length. A restraint similar to the brace can also be placed at the end points of a column [Fig. 1(d)]. It is known that when both ends are fixed, the buckling load increases approximately by twofold (see [63], p. 647).

While the compression between two nonpolar SCs is far different in material composition from the compression force in a column, their needing stabilization or steadying is the same. Furthermore, the equations that characterize the stabilization should be based on the same principle. In the next subsection we will derive the buckling load for a pair of nonpolar SCs in compression. It will show the extent that a lateral support can prevent the pair from slipping off their line of action. This is analogous to adding a brace to a column. With the multitude of nonpolar contacts in the interior, we can expect the numerous compressions coming from all directions to stabilize some structurally critical ones. This serves as the basis of our view that the thermodynamic stability of protein structure could be rooted in a form of mechanical stability.

B. Buckling load of a blocking interaction

We now describe some scenarios of the instability of the blocking action and derive the buckling load when the action is stabilized by a lateral support. In mechanics, a stable equilibrium is always associated with a mechanical energy minimum (e.g., see [70], p. 151). Thus, this derivation demonstrates that protein stability in our model is associated with some form of energy minimum.

As shown in Fig. 8 the instability in blocking can take many forms in abstract. In Figs. 8(a) and 8(b) the loss of blocking can be depicted by a single section, i.e., in two dimensions. In Fig. 8(c), the process has to be seen in three dimensions. In Figs. 8(a) and 8(c), the tips of two interacting SCs slip off one another much like the ball on a mound that would roll off (Sec. VIA 2). In Fig. 8(b), the tip of

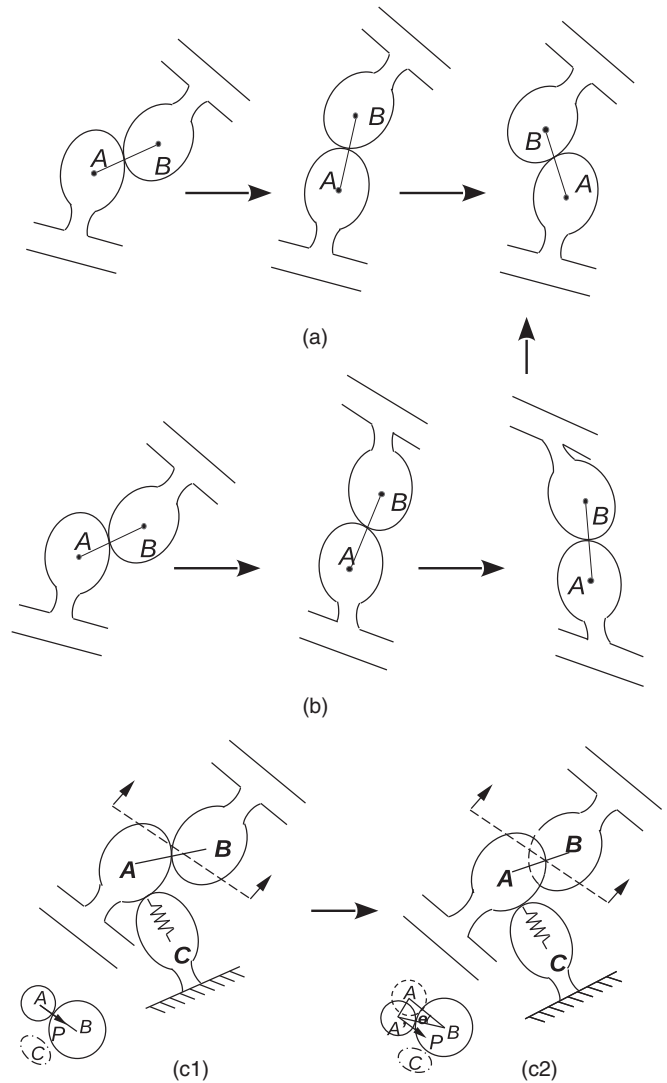


FIG. 8. Schematic drawings of a SC (a) slipping off from a strong blocking action line in 2D, (b) flipping due to a strong compression from the interacting SC in 2D, and (c) slipping off to a side of the interacting SC in 3D. (c) Introduces a third SC C that provides lateral support to A and adds a section between SCs A and B (indicated by two arrows and a dashed line) because the motion is in 3D. (c1) Shows two SCs in near perfect head-to-head compression. (c2) Shows that the two SCs have moved to each other’s side and the action line has moved towards the orthogonal direction with respect to the substructure axes. The sections are shown underneath. The original position of SC A in (c2) is marked with dashed line. The move of action lines is more clearly seen in the sections. In (c1) load P on A is originally balanced with blocking force $-f_{AB}$. In (c2), because the SC A has moved an angle of θ , P will be off the perfect action line by θ and will have a tangential component with respect to SC B . Here SC C pushes SC A from side as often observed in the tightly packed interior. Its action is like having a spring supporting SC A . SC C is also shown in the sections. It is in dashed lines because it is much lower and will not be on the section plane per se.

one SC has pushed the tip of the other so far that the latter flipped.

We can see how a lateral support resists a perturbation load from Fig. 8(c1). Here we assume the SC with centroid at A

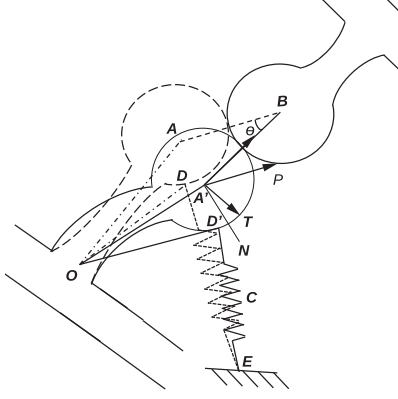


FIG. 9. Schematic drawing for deriving Eq. (18). This is simplified from Fig. 8(a). A lateral support from a third SC C is added and is represented by a spring (DE) acting at point D . The load P is assumed to be in the direction of AB along which the two SCs are originally in perfect nonslip compression. After a random move, point A moves to A' and there is a deviation by an angle of θ . $A'T$ will be the direction of the tangential component of load P . As the angle θ is small, it can be assumed that the two contacting surfaces are approximately two circles so that $|AB| = |A'B|$. Note that θ is drawn exaggerated in the illustration. More details are in the main text.

(referred as SC A) is under the load P . There is a minor static friction between SCs A and B . Furthermore, there is the lateral support, another compression from a third SC C . It pushes SC A from side. This combined with the friction will keep SC A from slipping off under normal load. This force from C can be modeled as a spring acting at around halfway of the major axis of the ellipsoid of SC A . The restoring force that the spring can generate is $k\Delta$, where k is the stiffness and Δ the deformation or displacement.

A formula can be derived for the *buckling load* P above which the lateral support starts to fail and a SC may slip off the compressive action position. For this we turn to Fig. 9 where we treat the 2D case of Fig. 8(a) but add a supporting SC C which is represented by its functional part, the spring. Also for simplification, we represent the ellipses by their respective osculating circles at their contacting points (i.e., the circles of the same radii of curvature). There is an applied load P on SC A . Assume P is exactly in line with AB . In their perfect acting position, there is no tangential force on SC A . So there is enough support for SC A to prevent slippage and the blocking should persist at a normal load.

Suppose there is a random move of SC A off the action line to A' , which can be viewed as a rotation of OA rotating to OA' . The load P will form a small angle θ with the perfect compression line, now $A'B$. This will produce a tangential component on SC A , $P \sin\theta$ in the direction of $A'T$. The spring of lateral support now will be compressed and will resist this move. When θ is small, all the other rotations will be small too. Assume OA rotates a small angle θ^* . OD then should rotate the same θ^* to OD' . Let $r = |OD| = |OD'|$. The displacement of the spring then is $\Delta = r\theta^*$. Let $R = |AB| = |A'B|$, $L = |OA| = |OA'|$. We then have the arc length $|AA'| = R\theta = L\theta^*$.

Thus, $\theta^* = \frac{R}{L}\theta$. R, L are both constant. Let $\rho = \frac{R}{L}$. We have $\Delta = \rho r\theta$. The restoring force will be $f = k\Delta = k\rho r\theta$.

The load P and the spring are acting at different positions of SC A . Thus, the equilibrium in moment is needed. Here the rotational center is at O for both load P and spring force f . The force arms are associated with $L = OA'$ and $r = OD'$, respectively, for P and f . We can draw a line $A'N$ in the figure, so that $A'N \perp OA'$. Denote the angle between $A'T$ and $A'N$ as $\angle(A'T, A'N)$. The moment for P then is

$$M_P = P \sin\theta \cos\angle(A'T, A'N)L.$$

It can be shown that $\angle(A'T, A'N) = \frac{\theta + \theta^*}{2}$.⁸ Since $\cos(\frac{\theta + \theta^*}{2}) \approx 1$ because both θ and θ^* are small, $M_P = P \sin\theta L$. This is to be balanced by the moment from the spring, which is $f \cos(\beta - \pi/2)r$ where β is the angle between the spring direction and the vector of DO or $D'O$, i.e., that between DE and DO (or between $D'E$ and $D'O$).

$$P \sin\theta L = \rho k r \theta \sin\beta r.$$

Because the rotation θ^* is small, β can be considered constant. Let $C = \rho \sin\beta$.

$$P \sin\theta L = C k r^2 \theta.$$

When θ is small, $\sin\theta = \theta$. Thus,

$$P = C k r^2 / L. \quad (18)$$

One may notice that this equation looks similar to Euler's buckling load [Eq. (17)]. Similar formulas are arrived at in many textbook examples of buckling ([63], p. 636, [64], p. 83, [70], p. 154). This derivation has to be substantially more complex, though, because we are dealing with the *stability of the equilibrium involving an interaction between two objects* than a single structural member like a column.

C. Stability and determinacy of a core assembly viewed through truss representation

The observation that each mutual blocking or double blocking has the potential of reducing one DOF can be viewed in the light of the truss representation of a core assembly. Here to simplify the discussion, the truss bars are assumed to carry both tension and compression and the interlocking is nonpolar. As introduced earlier (Sec. III A), in 3D the basic equation is $3j = b + r$. In general, just in terms of equation system solving, if we have $b + r$ number of unknowns, they will allow the $3j$ equations to have a unique solution.

However, there is a caveat: There are cases, even when the basic equation is satisfied, the truss is still unstable and its corresponding system of equations does not have a solution. This is most often when a certain number of reactions or bars are concurrent at a point or are parallel. As a result, their ability in

⁸ $\angle(A'T, A'N) = 2\pi - \angle OA'N - \angle BA'T - \angle OA'A - \angle BA'A$. But $\angle OA'N = \angle BA'T = \pi/2$. Because $|OA| = |OA'|$, $|BA| = |BA'|$, $\angle OA'A = (\pi - \theta^*)/2$, $\angle BA'A = (\pi - \theta)/2$. Thus, $\angle(A'T, A'N) = \frac{\theta + \theta^*}{2}$.

resisting a moment or force will be reduced. These are termed “improper constraints” [61] or “geometrically unstable” [62] in literature. This can also happen when a disproportional number of bars and reactions are overconstraining a part of the truss, thus leaving other parts underconstrained. By excluding these caveat conditions, satisfying the basic equation can be a *sufficient* (as well as necessary) condition for the truss to be stable. We note that the truss is stable *if and only if* its linear system of equations is solvable. Being algebraic, this criterion for stability is complete. Thus, the notion of exclusion above can and should be interpreted as applying this algebraic criterion.

The equation $3j = b + r$ implies that so far as the structure’s stability and determinacy is concerned, the relationship between b and r is irrelevant as long as their sum equals $3j$. But, internally for the truss structure, how many bars (b) or reactions (r) there are makes a difference. Bars are internal structural members. Too few of the bars, the structure is not standing on its own. (See “Internal stability” [61].) The truss as a solid body has six DOF. Thus, $r = 6$ will exactly provide the reaction support to fix the body in space. If we still have a stable and determinate structural solution, i.e., $3j = b + 6$, then the internal structural members have held the structure together, in an internally stable state. Let j_A and b_A be the joints and bars in the truss of core assembly A . The sufficient condition for the assembly A to be in that state then is

$$b_A = 3j_A - 6 \quad (19)$$

assuming each constituent substructure (represented in the truss) is itself in that state and assuming the exclusion of improper constraints and overconstraining mentioned above.

With this understanding, we check how the condition of Eq. (19) is satisfied when interlockings are added. Assume assembly A is generated from combining two assemblies, A_1 and A_2 , each satisfying the above requirement $3j_1 = b_1 + 6$ and $3j_2 = b_2 + 6$, where subscripts 1 and 2 indicate which of the assemblies the joints or bars belong to. With the combination but without counting the interlocking, the above two equations will be added on both sides:

$$3j = b + 12,$$

where $j = j_1 + j_2$, $b = b_1 + b_2$. Here, to satisfy Eq. (19), six reactions must be replaced by six bars. Clearly, two interlockings, with each being between one substructure from A_1 and another from A_2 , are exactly needed. This is quite understandable in terms of the change of DOF: when we combine A_1 and A_2 , we are essentially fixating A_1 relative to the body frame of A_2 (or vice versa). There are exactly six DOF to be removed. Evidently, six bars from two interlockings can do just that.

Let the number of substructures in assembly A be $s = |A| = |\{H_i\}|$ and the number of interlockings in A be n_I . As just explained, each substructure has extra six DOF to be removed when added to the assembly and it takes two interlockings to do that. For s substructures, this requires total of $2(s - 1)$ interlockings. When none of the interlockings are improperly constraining or overconstraining and when the reactions are just enough for fixing the assembly in space (i.e.,

$r = 6$), the assembly is uniquely fixated. That is,⁹

$$\begin{aligned} n_I &= 2(s - 1) \\ \Rightarrow \text{Assembly } A &\text{ has a stable and determinate structure.} \end{aligned} \quad (20)$$

Notice that since s is small, $n_I - (s - 1) = s - 1$, where $(s - 1)$ is the minimal number of interlockings for connecting the substructures, is quite small. For example, when $s = 4$ only $n_I = 6$ interlockings are enough to fixate the assembly fully in the space. $n_I = n_\Sigma + n_d$ are in Tables I (for n_Σ) and II (for n_d) (Sec. V) and in the tables for beta proteins (Supplemental Material [66]). Many PDB structures there can have enough redundant interlockings to meet the requirement of Eq. (20). Many more are only short of two or three interlockings. Given that the interlockings are calculated using a single threshold for chains with diverse sequence composition (Sec. IV D), this is quite satisfactory.

VII. CONCLUSION

This research is based on the premise that the protein structure, however small and intricate, is still a structure in mechanics sense. That is, when treated as a time and ensemble average, its structural members must be in a static equilibrium and the equilibrium must be stable. Based on the observation that the protein nonpolar interior is strongly repulsive, i.e., compressive in static terms, our static model infers that this force when positioned properly will be sufficient to restrict the relative motion of protein substructures and fixate the protein to a unique or distinct topology through the mechanism of interlocking. The following findings are obtained in developing this model: (1) Interlockings as defined in Eqs. (7) and (8), prevalent in protein interior, provide a mechanism for restricting relative motion between substructures. (2) The core assembly, while being built up from interlockings that are often deep in the interior, is able to hold the substructures fixated as indicated by Eq. (20). (3) The multitude of support reactions arising from nonpolar contacts enables even load distribution in 3D. (4) The lateral support for stabilizing the compressions in the interlocking can be supplied by nonpolar contacts [Eq. (18)]. In applying the model to analyzing sets of PDB structures and decoys we have shown that indeed native structures, depending on their sequence composition, have majorities of their substructures organized in interlocked positions at different levels of strength. The decoys either are unable to establish a sufficiently interlocked interior or contain destabilizing factors, e.g., buried unneutralized groups or multiple prolines inside helices, in an apparently interlocked interior.

ACKNOWLEDGMENT

This research is partially supported by NIH Grant No. GM 65628,

⁹In reality, the compression-only nature of the bars in the core assembly requires a multitude of support reactions (Sec. II B). Thus, practically only the “stable” part of the sufficient condition matters.

- [1] J. Richardson, Anatomy and taxonomy of protein structure, *Adv. Protein Chem.* **34**, 167 (1981).
- [2] J. Richardson and D. C. Richardson, Principles and patterns of protein conformation, in *Prediction of Protein Structure and the Principles of Protein Conformation*, edited by G. Fasman (Plenum, New York, 1989).
- [3] C. Chothia and A. V. Finkelstein, The classification and origin of protein folding patterns, *Annu. Rev. Biochem.* **59**, 1007 (1990).
- [4] F. R. Salemme, Conformational and geometrical properties of beta-sheets in proteins, *J. Mol. Biol.* **146**, 143 (1981).
- [5] A. G. Murzin and A. V. Finkelstein, General architecture of the alpha-helical globule, *J. Mol. Biol.* **204**, 749 (1988).
- [6] J. U. Bowie, Helix packing angle preferences, *Nat. Struct. Biol.* **4**, 915 (1997).
- [7] C. Chothia and J. Janin, Relative orientation of close-packed β -pleated sheets in proteins, *Proc. Natl. Acad. Sci. USA* **78**, 4146 (1981).
- [8] C. Chothia, Principles that determine the structure of proteins, *Annu. Rev. Biochem.* **53**, 537 (1984).
- [9] N. L. Harris, S. R. Presnell, and F. E. Cohen, Four helix bundle diversity in globular proteins, *J. Mol. Biol.* **236**, 1356 (1994).
- [10] J. Janin and C. Chothia, Packing of α -helices onto β -pleated sheets and the anatomy of α/β proteins, *J. Mol. Biol.* **143**, 95 (1980).
- [11] A. M. Lesk, C. Branden, and C. Chothia, Structural principles of alpha/beta barrel proteins: The packing of the interior of the sheet, *Proteins* **5**, 139 (1989).
- [12] F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor, Analysis of the tertiary structure of protein β -sheet sandwiches, *J. Mol. Biol.* **148**, 253 (1981).
- [13] E. G. Hutchinson, R. B. Sessions, and J. M. Thornton, Determinants of strand register in antiparallel β -sheets of proteins, *Protein Sci.* **7**, 2287 (1998).
- [14] M. Edwards, M. J. E. Sternberg, and J. M. Thornton, Structural and sequence patterns in the loops of $\beta\alpha\beta$ units, *Protein Eng. Des. Sel.* **1**, 173 (1987).
- [15] C. Tanford, Protein denaturation: Part c. Theoretical models for the mechanism of denaturation, *Advances in Protein Chemistry* (Elsevier, Amsterdam, 1970).
- [16] K. A. Dill, Dominant forces in protein folding, *Biochemistry* **29**, 7133 (1990).
- [17] N. Kurochkina and G. Privalov, Heterogeneity of packing: Structural approach, *Protein Sci.* **7**, 897 (1998).
- [18] K. A. Dill, T. M. Truskett, V. Vlachy, and B. Hribar-Lee, Modeling water, the hydrophobic effect, and ion solvation, *Annu. Rev. Biophys. Biomol. Struct.* **34**, 173 (2005).
- [19] C.-J. Tsai, J. V. Maizel Jr, and R. Nussinov, The hydrophobic effect: A new insight from cold denaturation and a two-state water structure, *Crit. Rev. Biochem. Mol. Biol.* **37**, 55 (2002).
- [20] A. Nicholls, K. A. Sharp, and B. Honig, Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons, *Proteins: Structure, Function, and Genetics* **11**, 281 (1991).
- [21] Y. Levy and J. N. Onuchic, Water mediation in protein folding and molecular recognition, *Annu. Rev. Biophys. Biomol. Struct.* **35**, 389 (2006).
- [22] M. Charton and B. I. Charton, The structural dependence of amino acid hydrophobicity parameters, *J. Theor. Biol.* **99**, 629 (1982).
- [23] R. Godawat, S. N. Jamadagni, and S. Garde, Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations, *Proc. Natl. Acad. Sci. USA* **106**, 15119 (2009).
- [24] M. Levitt and A. Warshel, Computer simulation of protein folding, *Nature (London)* **253**, 694 (1975).
- [25] T. Lazaridis and M. Karplus, Discrimination of the native from misfolded protein models with an energy function including implicit solvation, *J. Mol. Biol.* **288**, 477 (1999).
- [26] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes, Optimizing physical energy functions for protein folding, *Proteins* **54**, 88 (2004).
- [27] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J. Mol. Biol.* **268**, 209 (1997).
- [28] P. A. Bates, R. M. Jackson, and M. J. E. Sternberg, Model building by comparison: A combination of expert knowledge and computer automation, *Proteins* **29**, 59 (1997).
- [29] J. Pillardy, C. Czaplowski, A. Liwo, J. Lee, D. R. Ripoll, R. Kamierkiewicz, S. Odziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye, and H. A. Scheraga, Recent improvements in prediction of protein structure by global optimization of a potential energy function, *Proc. Natl. Acad. Sci. USA* **98**, 2329 (2001).
- [30] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, Ab initio construction of protein tertiary structures using a hierarchical approach, *J. Mol. Biol.* **300**, 171 (2000).
- [31] A. Fiser, R. Kinh Gia Do, and A. Sali, Modeling of loops in protein structures, *Protein Sci.* **9**, 1753 (2000).
- [32] A. Jaramillo, L. Wernisch, S. Hery, and S. J. Wodak, Folding free energy function selects native-like protein sequences in the core but not on the surface, *Proc. Natl. Acad. Sci. USA* **99**, 13554 (2002).
- [33] J. O. Wrabl, S. A. Larson, and V. J. Hilser, Thermodynamic propensities of amino acids in the native state ensemble: Implications for fold recognition, *Protein Sci.* **10**, 1032 (2001).
- [34] M. E. Milla, B. M. Brown, and R. T. Sauer, Protein stability effects of a complete set of alanine substitutions in arc repressor, *Nat. Struct. Biol.* **1**, 518 (1994).
- [35] B. M. Brown, M. E. Milla, T. L. Smith, and R. T. Sauer, Scanning mutagenesis of the arc repressor as a functional probe of operator recognition, *Nat. Struct. Biol.* **1**, 164 (1994).
- [36] W. A. Baase, L. Liu, D. E. Tronrud, and B. W. Matthew, Lessons from the lysozyme of phage t4, *Protein Sci.* **19**, 631 (2010).
- [37] W. A. Lim and R. T. Sauer, Alternative packing arrangements in the hydrophobic core of lambda repressor, *Nature (London)* **339**, 31 (1989).
- [38] A. M. Bonvin, H. Vis, J. N. Breg, M. J. Burgering, R. Boelens, and R. Kaptein, Nuclear magnetic resonance solution structure of the arc repressor using relaxation matrix calculations, *J. Mol. Biol.* **236**, 328 (1994).
- [39] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, Three key residues form a critical contact network in a protein folding transition state, *Nature (London)* **409**, 641 (2001).
- [40] B. V. B. Reddy, W. W. Li, I. N. Shindyalov, and P. E. Bourne, Conserved key amino acid positions (ckaaps) derived from the analysis of common substructures in proteins, *Proteins* **42**, 148 (2001).

- [41] M. D. Collins, M. L. Quillin, G. Hummer, B. W. Matthews, and S. M. Gruner, Structural rigidity of a large cavity-containing protein revealed by high-pressure crystallography, *J. Mol. Biol.* **367**, 752 (2007).
- [42] R. Samudrala and M. Levitt, Decoys R us: A database of incorrect conformations to improve protein structure, *Protein Sci.* **9**, 1399 (2000).
- [43] B. H. Park and M. Levitt, Complexity and accuracy of discrete state models of protein structure, *J. Mol. Biol.* **249**, 493 (1995).
- [44] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker, An improved protein decoy set for testing energy functions for protein structure prediction, *Proteins* **53**, 76 (2003).
- [45] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, The i-tasser suite: Protein structure and function prediction, *Nat. Methods* **12**, 7 (2015).
- [46] A. Kessel and N. Ben-Tal, *Introduction to Proteins* (CRC Press, Boca Raton, FL, 2018).
- [47] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, Charmm: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* **4**, 187 (1983).
- [48] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson *et al.*, *AMBER 2018* (University of California, San Francisco, 2018).
- [49] M. J. Abraham, T. Murtola, R. Schulz, and J. C. Smith, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX* **1-2**, 19 (2015).
- [50] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc.* **118**, 11225 (1996).
- [51] R. Samudrala, Y. Xia, M. Levitt, and E. S. Huang, A combined approach for ab initio construction of low resolution protein tertiary structures from sequence, *Pacific Symposium on Bio-computing* (International Society for Computational Biology, Leesburg, VA, 1999).
- [52] Charlotte M. Deane, Quentin Kaas, and Tom L. Blundell, Score: Predicting the core of protein models, *Bioinformatics* **17**, 541 (2001).
- [53] A. Kolinski, M. R. Betancourt, D. Kihara, P. Rotkiewicz, and J. Skolnick, Generalized comparative modeling (gencomp): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement, *Proteins* **44**, 133 (2001).
- [54] K. Yue and K. A. Dill, Folding proteins with a simple energy function and extensive conformational searching, *Protein Sci.* **5**, 254 (1996).
- [55] K. Yue and K. A. Dill, Constraint-based assembly of tertiary protein structures from secondary structure elements, *Protein Sci.* **9**, 1935 (2000).
- [56] Michael Levitt, Wikipedia, https://en.wikipedia.org/wiki/Michael_Levitt
- [57] P. Koehl and M. Levitt, A brighter future for protein structure prediction, *Nat. Struct. Biol.* **6**, 108 (1999).
- [58] D. P. Goldenberg and T. E. Creighton, Energetics of protein structure and folding, *Biopolymers* **24**, 167 (1985).
- [59] J. P. Staley and P. S. Kim, Complete folding of bovine pancreatic trypsin inhibitor with only a single disulfide bond, *Proc. Natl. Acad. Sci. USA* **89**, 1519 (1992).
- [60] M. Karplus and D. L. Weaver, Protein folding dynamics: The diffusion-collision model and experimental data, *Protein Sci.* **3**, 650 (1994).
- [61] R. C. Hibbeler, *Structural Analysis* (Pearson, New York, 2017).
- [62] K. Leet, C.-M. Uang, and J. Lanning, *Fundamentals of Structural Analysis* (McGraw-Hill, New York, 2020).
- [63] R. Craig and E. Taleff, *Mechanics of Materials* (Wiley, Hoboken, NJ, 2017).
- [64] S. P. Timoshenko and J. M. Gere, *Theory of Elastic Stability* (Dover, New York, 2009).
- [65] G. J. Simitses, D. Guggenheim, and D. H. Hodges, *Fundamentals of Structural Stability* (Elsevier, Amsterdam, 2006).
- [66] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.106.024410> for data and discussions on compressive support, substructure gaps, instability of blocking, resolving truss indeterminacy, core assembly comparison, and related topics.
- [67] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York, 1991).
- [68] Wikipedia, π stacking, <http://en.wikipedia.org/wiki/pistacking>
- [69] B. Lee and Richards FM, The interpretation of protein structures: Estimation of static accessibility, *J. Mol. Biol.* **55**, 379 (1971).
- [70] J. P. Den Hartog, *Mechanics* (Dover, New York, 2013).
- [71] D. L. Schodek, *Structures* (Pearson, New York, 2004).
- [72] S. Kumar and R. Nussinov, Salt bridge stability in monomeric proteins, *J. Mol. Biol.* **293**, 1241 (1999).
- [73] W. Humphrey, A. Dalke, and K. Schulten, VMD – visual molecular dynamics, *J. Mol. Graphics* **14**, 33 (1996).
- [74] P. Andersen, *Statically Indeterminate Structures: Their Analysis and Design* (Ronald Press, New York, 1953).
- [75] K. Yue and K. A. Dill, Sequence structure relationship of proteins and copolymers, *Phys. Rev.* **48**, 2267 (1993).
- [76] K. Yue, K. Fiebig, H. S. Chan, P. D. Thomas, H. S. Chan, E. Shakhnovich, and K. A. Dill, A test of lattice protein folding algorithms, *Proc. Natl. Acad. Sci. USA* **92**, 325 (1995).
- [77] Wikipedia, Steel bank common lisp, <http://en.wikipedia.org/wiki/sbcl>
- [78] Maxima project at SourceForge, Maxima manual, <https://maxima.sourceforge.io/docs>
- [79] R. A. Laskowski, J. A. Rullmannn, M. W. MacArthur, R. Kaptein, and J. M. Thornton, Aqua and procheck-nmr: Programs for checking the quality of protein structures solved by nmr, *J. Biomol. NMR* **8**, 477 (1996).
- [80] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Funnel, pathways, and the energy landscape of protein folding: A synthesis, *Proteins* **21**, 167 (1995).
- [81] A. R. Fersht, Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism, *Proc. Natl. Acad. Sci. USA* **97**, 1525 (2000).
- [82] K. A. Dill and H. S. Chan, From levinthal to pathways to funnels, *Nat. Struct. Biol.* **4**, 10 (1997).
- [83] J. L. MacCallum, M. Sabaye Moghaddam, H. S. Chan, and D. P. Tieleman, Hydrophobic association of alpha-helices, steric dewetting and enthalpic barriers to protein folding, *Proc. Natl. Acad. Sci. USA* **104**, 6206 (2007).

- [84] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, A backbone-based theory of protein folding, *Proc. Natl. Acad. Sci. USA* **103**, 16623 (2006).
- [85] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, Forces contributing to the conformational stability of proteins, *FASEB J.* **10**, 75 (1996).
- [86] C. F. Lopez, R. K. Darst, and P. J. Rossky, Mechanistic elements of protein cold denaturation, *J. Phys. Chem. B* **112**, 5961 (2008).
- [87] P. J. Fleming and F. M. Richards, Protein packing: Dependence on protein size, secondary structure and amino acid composition, *J. Mol. Biol.* **299**, 487 (2000).
- [88] T. Rivlin, *An Introduction to Approximation of Functions* (Dover, New York, 1969).
- [89] K. Yue and K. A. Dill, Forces of tertiary structural organization of globular proteins, *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).
- [90] C. A. Coulomb, Essai sur une application des regles de maximis & minimis a quelques problemes de statique, relatifs a l'architecture, in *Memoires de Mathematique & de Physique*.