Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: A consensus

Nicolas E. G. Buchler and Richard A. Goldstein[a)]
Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109-1055

(Received 8 October 1999; accepted 8 November 1999)

A variety of analytical and computational models have been proposed to answer the question of why some protein structures are more ''designable'' (i.e., have more sequences folding into them) than others. One class of analytical and statistical-mechanical models has approached the designability problem from a thermodynamic viewpoint. These models highlighted specific structural features important for increased designability. Furthermore, designability was shown to be inherently related to thermodynamically relevant energetic measures of protein folding, such as the foldability $\mathcal{F}$ and energy gap Δ_{10}. However, many of these models have been done within a very narrow focus: Namely, pair–contact interactions and two-letter amino-acid alphabets. Recently, two-letter amino-acid alphabets for pair–contact models have been shown to contain designability artifacts which disappear for larger-letter amino-acid alphabets. In addition, a solvation model was demonstrated to give identical designability results to previous two-letter amino-acid alphabet pair–contact models. In light of these discordant results, this report synthesizes a broad consensus regarding the relationship between specific structural features, foldability $\mathcal{F}$, energy gap Δ_{10}, and structure designability for different energy models (pair–contact vs solvation) across a wide range of amino-acid alphabets. We also propose a novel measure Z_d^k which is shown to be well correlated to designability. Finally, we conclusively demonstrate that two-letter amino-acid alphabets for pair–contact models appear to be solvation models in disguise. © *2000 American Institute of Physics.* [S0021-9606(00)52305-8]

INTRODUCTION

A significant over-representation of certain protein folds in the biological database has been extensively documented.[1–3] After uncoupling evolutionary and functional relationships between protein families in the database, it was noticed that certain structural ''superfolds'' such as α/β doubly wound and triosephosphate isomerase (TIM) barrels are represented ~11–13 times compared to, for instance, the jelly roll motif which has been observed only three times.[2] The possible origins of this dispersion in structural representation in the biological database poses an interesting question. Are there energetic, kinetic, or topological reasons for this dispersion? Concerning this question, Finkelstein and co-workers were the first to use simple analytical models for protein sequence energy landscapes to demonstrate that energetic and topological constraints can indeed stabilize and lead to easier ''design'' of certain protein motifs by random sequences.[4–8] This mutually reinforcing relationship between protein structure designability and energetic stabilization has become the focus of recent analytical and computational work on lattice proteins.

One theoretical model that attempts to address kinetic and stability issues of protein folding and structure designability was based on ideas first developed by Bryngelson and Wolynes. Two thermodynamic transitions are considered possible in protein folding: One to the folded, native state at a temperature T_f and the other to a glassy state at a temperature T_g.[9–14] For temperatures below T_g, the liquidlike protein chain entropy drops to zero (in the thermodynamic limit) and the chain becomes solidlike and ''frozen'' in any one of its low-energy, metastable states. Consequently, as the temperature approaches T_g, folding kinetics become slow and it becomes difficult for the protein to transit from misfolded local minima to other stable states. T_f, on the other hand, defines the temperature at which the free energy of the folded state is deep enough to be preferentially populated over other kinetically accessible conformations and be stable to thermal fluctuations. Thus, assuming that protein sequence foldability requires an equilibrium temperature regime which is both adequately below T_f for the folded state to be stable yet sufficiently above T_g for the folded state to be accessible, Wolynes and co-workers postulated that optimal folding energy landscapes would seek to maximize T_f and minimize T_g or, equivalently, increase the ratio T_f/T_g.[15,16] Using the random energy model (REM), it was shown that this ratio is equal to

$$\frac{T_f}{T_g} = \sqrt{\frac{\mathcal{F}^2}{2S_0}} + \sqrt{\frac{\mathcal{F}^2}{2S_0} - 1}, \quad \text{where } \mathcal{F} = \frac{\langle E \rangle - E^k}{\sigma}, \quad (1)$$

where $\langle E \rangle$ is the average energy of the protein chain in all conformations, E^k is the global energy minimum belonging to the native structure k, and σ describes the variance and ''roughness'' of this REM energy landscape. Clearly, T_f/T_g is a monotonically increasing function of $\mathcal{F}$ (equivalently

[a)]Author to whom correspondence should be addressed; also at: Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055; telephone: 734-763-8013; fax: 734-647-4865; electronic mail: richardg@umich.edu

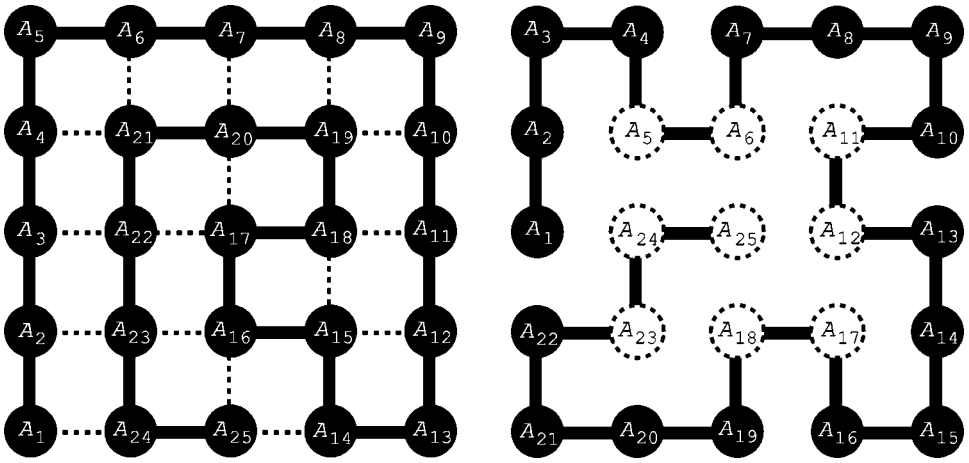

FIG. 1. On the left is a generic sequence in the most designable 5×5 lattice structure for the *pair–contact* model calculated by Monte Carlo sampling. On the right is a generic sequence in the most designable 5×5 lattice structure for the *solvation* model, also calculated by Monte Carlo sampling. The interactions in each energy model are represented by dotted lines and, for maximally compact lattice proteins, are constant across all structures (there are always 16 pair–contacts, nine buried residues per structure). Note the large number of long-range pair–contacts between amino-acids far from one another in the sequence in the *pair–contact* model. For the *solvation* model, in contrast, there is a high degree of symmetry and regularity in the buried–exposed pattern of the most designable structures. Both patterns in these highly designable structures for different models can be explained by the same mechanism. Namely, all these structural interactions are rare with respect to the structural ensemble.

known as a *Z*-score in the protein design literature[17,18]), which was termed the sequence ''foldability.'' While the REM ignores all correlations in the free-energy landscape, it has been shown that maximizing the foldability for protein sequences results in a reduction in the depth of metastable traps and the stabilization of conformations similar to the native state producing the funnel-like energy landscapes central to good folders.[19–22] Additionally, it was verified both with molecular dynamic and Monte Carlo kinetic simulations that faster folders had higher average foldabilities and larger T_f/T_g.[15,16,23,24]

A different energy landscape measure, Δ_{10}, also related to folding kinetics and thermodynamic stability was first introduced by Shakhnovich and co-workers.[25,26] This energy gap is defined as the difference in energy between the global-energy minimum native state E^k and that of the next highest energy E^1. Upon running Monte Carlo kinetic simulations of lattice proteins and analyzing their energy landscapes, Shakhnovich and co-workers noticed that the relevant statistical feature for fast folding was a large energy gap Δ_{10}.[27,28] Recent work using the random energy model and lattice proteins demonstrates that in spite of their obvious differences, $\mathcal{F}$ and Δ_{10} are inherently correlated.[29] Hence, for all intensive purposes, $\mathcal{F}$ and Δ_{10} can be used interchangeably as similar measures of fast-folding and native state stability.

The existence of a simple statistical measure of an energy landscape, well-correlated with increased thermodynamic stability and the ability of the protein to fold, prompted Govindarajan and Goldstein to use foldability $\mathcal{F}$ to analyze structure designability. Upon calculating the maximum-possible ''optimal'' foldability F^k_{opt} within a pair-contact model for a prototypical sequence folding into a target structure *k*, Govindarajan and Goldstein demonstrated that there was a dispersion in possible F^k_{opt} across all structures.[30] They postulated that those structures with the largest F_{opt} would be the most designable, as there would be many possible sequences far from the optimum which are still adequately foldable.[30,31] Conversely, a protein structure with low F^k_{opt} would be poorly designable, as only the relatively rare sequences with close-to-optimal interactions would be able to successfully fold into these structures. Additionally, for any relevant value of $\mathcal{F}^\star$, large F^k_{opt} structures would have a larger volume of foldable sequences satisfying $\mathcal{F}>\mathcal{F}^\star$ than low F^k_{opt} structures. Hence, an implicit conclusion of this foldability–designability model is that those sequences folding into highly designable structures should also have larger, average foldability $\langle\mathcal{F}\rangle$ and energy gap $\langle\Delta_{10}\rangle$ than their lowly designable counterparts. Using a statistical model of protein interaction space, Govindarajan and Goldstein analytically demonstrated a strong, positive correlation between F^k_{opt} and structure designability V_k (i.e., the volume of sequence space folding into structure *k*).[31] The appeal of having a measure F^k_{opt} well correlated to V_k is that it can be calculated *a priori* using only structural information. In addition, it is easily interpretable which structural features lead to higher F^k_{opt} and, hence, designability. Govindarajan and Goldstein explicitly showed that those structures with ''rare'' or ''uncommon'' pair–contacts with respect to the ensemble of all possible structures had larger F^k_{opt}. Even with local, secondarylike propensities added to the lattice protein model, the optimal foldability was predominantly determined by those rare, nonlocal pair–contacts between distant sequence positions.[32] Hence, one would expect highly designable pair–contact lattice proteins to have a maximum number of long-range pair–contacts [see Fig. 1(a)].

In a computational tour de force, Li and co-workers exhaustively enumerated the structural designability of two-letter amino-acid alphabet hydrophobic polar (HP) sequences in different lattice geometries [6×62D (two dimensional) and $3\times3\times33$D (three dimensional)] for a pair–contact model.[33] Not only was it consistently found that some struc-

tures were more designable than others, but larger V_k structures also had bigger $\langle \Delta_{10} \rangle$ compared to lowly designable structures. Of particular mention, there was a large, discontinuous jump in $\langle \Delta_{10} \rangle$ between highly designable and lowly designable structures. Similar in logic to work by Bryngelson and Pande,[34,35] Tang and co-workers posited that these highly designable structures were inherently more stable to perturbation and/or mutation.[34,35] A recent deluge of computational lattice protein designability work makes a strong case for the inherent relationship between thermodynamic and mutational stability, faster-folding kinetics, extensive neutral networks, and the large designability of particular structures.[36–46] All in all, these computational results voice support for the foldability–designability model of Govindarajan and Goldstein. However, in contrast to the results of Govindarajan and Goldstein, based on the structural features of highly designable structures calculated from their HP amino-acid alphabet Tang and co-workers concluded that ''proteinlike'' symmetries increased designability rather than a maximum of long-range pair–contacts [see Fig. 1(b)].

There are several reasons to suspect the existence of amino-acid alphabet dependent artifacts endemic to simple HP amino-acid codes; artifacts which should disappear for larger amino-acid alphabets.[47] Previous work has shown that when m_{eff} (the effective entropy of sequence space) and γ (the conformational entropy per residue) satisfies $m_{\mathrm{eff}} < \gamma$, there exists a sizable number of sequences with degenerate global-energy minima which are considered unfoldable.[6,48,49] However for larger amino-acid alphabets where $m_{\mathrm{eff}} > \gamma$, such ground-state degeneracies rapidly vanish. Consequently, specific designability conclusions for two-letter alphabets might not be valid for those higher-letter alphabets which have more diverse pair–contact interactions and larger m_{eff}. Recent work of ours specifically explored the effect of amino-acid alphabet size on lattice protein structure designability for a pair–contact model.[50] Indeed, it was noticed that small amino-acid alphabets, such as the two-letter HP and Li *et al.* amino-acid code, had substantially different designability results from those of larger amino-acid alphabets, such as the Miyazawa–Jernigan (MJ) 20-letter code or the independent interaction model. Namely, those structures which were highly designable in two-letter amino-acid alphabets had mediocre designabilities with large-letter alphabets and vice versa. Moreover, with respect to the role of m_{eff} and γ, it was shown that the large number of degenerate ground states for smaller amino-acid alphabets was not the source of designability differences. In short, amino-acid alphabet dependent artifacts are inherent to the size of the alphabet rather than the details of amino-acid pair–contact interactions or the abundance of degenerate ground states.

An intriguing theoretical paper by Li and co-workers, based on a solvation model of lattice proteins where the energetics are only dependent on those residues buried in the protein core, elegantly recast the designability problem into a vector framework.[51] Namely, the energy of a sequence in a particular structure was equal to the Euclidean distance between the solvation structure k and the sequence vector S. Thus, a sequence S folds into the solvation structure k which is closest in distance to it. By numerically estimating the volume of sequence space V_k folding into a solvation structure k, Li and co-workers showed that those structures, which had a minimal density of other structures in their vicinity, were the most designable. In summary, Li and co-workers reasoned that those ''atypical'' or rare structures residing far away from the other solvation structures; hence having a low density of surrounding structures, were the most designable. Further analysis demonstrated that highly designable structures in a solvation model were identical to the ones they calculated for a pair–contact model with an HP two-letter amino-acid alphabet: That is, structures with lots of ''proteinlike'' symmetries.

Reconciling these results of Li and co-workers with our previous conclusions regarding alphabet-size artifacts and structural features (i.e., long-range pair–contacts) important for large designability, it is tempting to speculate that perhaps the two-letter pair–contact model is a solvation model in disguise. Indeed, recent papers by Ejtehadi and co-workers delineate a mathematical framework that highlights such a link between solvation and pair–contact models for two-letter HP amino-acid alphabets.[52,53] Namely, they cleverly begin with a special case with mixing coefficient $\gamma_M = 0$, that has pair–contact interactions of the form $\mathrm{HH} = -2\epsilon - \gamma_M$, $\mathrm{HP} = -\epsilon$, $\mathrm{PP} = 0$. This special case ($\gamma_M = 0$) allows these possible pair–contact interaction energies $\gamma(\mathcal{A}_i, \mathcal{A}_j)$ of two-letter amino-acid alphabets to be rewritten as a linear sum

$$\gamma(\mathcal{A}_i, \mathcal{A}_j) = \sigma_{\mathcal{A}_i} + \sigma_{\mathcal{A}_j} \quad \text{where} \quad \sigma_{\mathrm{H}} = -\epsilon \quad \text{and} \quad \sigma_{\mathrm{P}} = 0, \qquad (2)$$

where $\sigma_{\mathcal{A}_i}$ is the energy of having amino-acid $\mathcal{A}_i$ form *any* pair–contact, regardless of the identity of the other amino-acid. In this special linear interaction framework, the energy of a particular sequence in a structure becomes not a question of *which* amino-acid types form pair–contacts with one another, but rather *how many* pair–contacts does a particular amino-acid form in any given structure. Herein lies the connection between the pair–contact and solvation model, as those sequence positions buried in the protein core inherently form more pair–contacts than those on the surface.

For this linear interaction framework, there exists only a small subset of the total pair–contact structures that have a unique interaction pattern. Thus, most structures will have degenerate global-energy minima and remain completely undesignable for this two-letter linear interaction model ($\gamma_M = 0$). Only by having a nonzero mixing parameter γ_M, which energetically favors hydrophobic pair–contacts, can one split the ground-state degeneracy of these undesignable structures. As the Li *et al.* two-letter HP code is but a small, mixing perturbation ($\gamma_M = 0.3$) of the ''solvationlike'' linear interaction model, it should not be surprising that they have pair–contact designability results similar to that of a solvation model. In addition, Ejtehadi and co-workers provide a quantitative explanation for the observed discontinous jump in $\langle \Delta_{10} \rangle$ between the highly designable and the lowly designable structures (previously undesignable for $\gamma_M = 0$).[33,52,53] Additional work has shown that some pair–contact structures can remain stable and highly designable across an entire range of mixing parameters γ_M for any two-letter amino-acid alphabet.[54,55]

More recently, an analytical, statistical mechanical treatment for calculating the volume of two-letter HP sequence space folding into particular 2D pair–contact structures has been presented by Kussell and Shakhnovich.[56] Upon decomposing structures into various strings of pair–contacts, which in 2D can form either strands or loops, Kussell and Shakhnovich established that highly designable 2D structures for two-letter amino-acid alphabets should have the following properties: (1) No loops, (2) a maximum number of two-length strands, and (3) a minimum number of larger-length strands.

Given the wide range of these results, this report is an attempt to synthesize a broad consensus regarding the connection between sequence energetics and structure designability for different energy models and across a wide range of amino-acid alphabets. The latter is of specific interest as there is evidence of designability artifacts for smaller amino-acid alphabet sizes, yet most research has considered only two-letter amino-acid alphabets in pair–contact models. Thus, our questions are: How do the aforementioned energetic measures $\mathcal{F}^k_{\rm opt}$, $\langle\mathcal{F}\rangle$, and $\langle\Delta_{10}\rangle$ correlate to designability V_k? Does their relationship break down when done with smaller amino-acid alphabets? Within the solvation model?

We begin by introducing a universal framework abstracted from specific energy models and amino-acid composition with which to discuss principles of energetics and designability. Within this simple framework, the relationship between energy gap Δ_{10}, foldability $\mathcal{F}$, and designability V_k becomes geometrically interpretable. Furthermore, a novel measure Z^k_d, constructed to be correlated to designability, is introduced. We then calculate structure designability V_k across a range of amino-acid alphabets from two-letter to ''infinite''-letter (Monte Carlo sampling) codes in both solvation and pair–contact lattice models. It is established that V_k is indeed positively correlated to $\langle\mathcal{F}\rangle$, $\langle\Delta_{10}\rangle$, $\mathcal{F}^k_{\rm opt}$, and Z^k_d in both solvation and pair–contact energy models. In addition, for comparison to work by Kussell and Shakhnovich, we calculate the number of pair–contact loops and interaction strands versus $\langle V_k\rangle$. Several notable results stand out: (1) Across all energy models, the positive correlation between $\langle\mathcal{F}\rangle$, $\langle\Delta_{10}\rangle$, $\mathcal{F}^k_{\rm opt}$, Z^k_d, and V_k consistently deteriorates for two-letter amino-acid alphabets, (2) the 20-letter amino-acid alphabet is identical to Monte Carlo sampling; both alphabets accurately reflect the principles and structure designability of the underlying energy model, (3) two-letter amino-acid alphabets for a pair–contact model and the solvation model share identical highly designable structures (i.e., proteinlike and symmetric), and (4) while loops are negatively correlated to V_k across all amino-acid alphabets, for two-letter amino-acid alphabets having a maximum number of one-length and three-length strands, not two-length strands, leads to higher structure designability.

UNIVERSAL MODEL

For any given energy model, one commonly used framework for analyzing the designability and energetics of structures and sequences across various amino-acid alphabets has been the lattice protein. Lattice proteins are coarse-grained versions of proteins, where the level of detail focuses on amino-acids as entities occupying lattice points and protein conformations as self-avoiding walks on these regular lattices [see Figs. 1(a) and 1(b)]. Clearly, this ignores some very real aspects of proteins, such as atoms, backbone angles, sidechain packing, etc. Nevertheless, lattice proteins have a rich history in theoretical biophysics, not only because their simplicity captures salient features of biopolymers such as excluded volume and topology, but because the number of conformations is finite and amenable to analysis.[57] Throughout this paper, we restrict ourselves to maximally compact 5×5 2D lattice proteins. Our choice of this particular lattice model is threefold: (1) Maximally compact because in naturally occurring globular proteins, hydrophobic collapse is a dominant force and native structures are compact and solvent exclusive, (2) 2D because in a solvent framework, small 2D lattice conformations have a more realistic buried–exposed ratio, and (3) 5×5 because the number of possible conformations is large enough to be interesting, yet small enough to be feasible and have self-averaging statistics for designability calculations. Although there is a question whether 2D proteins exhibit problems in terms of cooperative folding kinetics,[58,59] energetic measures such as $\mathcal{F}$, Δ_{10} and designability V_k, as calculated in this report and others, are not dependent on the kinetic connectivity between structures. For maximally compact 5×5 2D lattice proteins, there are a total of 1081 unique structures after neglecting conformations related to one another by rotational or mirror symmetries. Thus, given such a set of lattice conformations, the energy landscape of a protein is determined by its amino-acid sequence and its peculiar energy model. In this report, we shall analyze these lattice proteins across two different energy models: The pair–contact and the solvation model.

In the pair–contact model of an N amino-acid length protein, as shown in Fig. 1(a) the energy of sequence S in conformation k is determined by the nonsequential, nearest-neighbor amino-acid pair–contacts that are formed. Mathematically, this is expressed as a linear equation:

$$E^k_S=\sum_{i<j}^{N,N}\gamma_S(\mathcal{A}_i,\mathcal{A}_j)\mathcal{D}^k_{i,j}=\boldsymbol{\gamma}_S\cdot\boldsymbol{\mathcal{D}}^k, \tag{3}$$

where $\gamma_S(\mathcal{A}_i,\mathcal{A}_j)$ is the pair–contact interaction energy of two arbitrary amino-acids whose values are specified in the definition of the amino-acid alphabet. $\mathcal{D}^k_{i,j}$ is shorthand for $\delta(\|\mathbf{r_i^k}-\mathbf{r_j^k}\|-a)$, where r^k_i, r^k_j are the sequence positions of amino-acids i, j in conformation k and a is the lattice spacing. In other words, $\mathcal{D}^k_{i,j}$ is equal to one if a pair–contact between sequence positions i and j is formed for structure k and 0, if otherwise. For the maximally compact 5×5 2D lattice protein, there are 132 such possible pair–contacts [132 is the dimension d of the vector space in Eq. (3)] of which 16 are actually formed for each structure k. Since each set of pair contacts for each of the 1081 possible structures is unique and nondegenerate, no vector $\mathcal{D}^k$ is identical to the other. Note that all $\mathbf{D}^k$ lie on a d-dimensional Euclidean hypercube with a fixed distance from the origin.

In the solvation model of an N amino-acid length protein,[51] as shown in Fig. 1(b), the energy of amino-acid

sequence S in conformation k is similarly expressed as

$$E_s^k = \sum_i^N \gamma_S(\mathcal{A}_i)\mathcal{D}_i^k = \boldsymbol{\gamma}_\mathbf{S} \cdot \mathbf{D}^k, \tag{4}$$

where $\gamma_S(\mathcal{A}_i)$ is the solvation energy of having amino-acid i of sequence S buried in the protein core rather than the solvent-exposed protein surface. $\mathcal{D}_i^k$ is equal to 1, if sequence position i is buried in structure k, and 0 otherwise. For the maximally compact 5×5 2D lattice protein, there are 25 possible residues that can be buried [25 is the dimension d of the vector space in Eq. (4)] of which nine are actually buried for each structure k. However, unlike the pair–contact model, there are degeneracies in the $\mathbf{D}^k$ vector; that is, there are some structures which are different from one another in the pair–contact model yet have identical solvation patterns in the solvation model. Thus, regardless of $\boldsymbol{\gamma}_\mathbf{S}$, the solvation model is unable to distinguish energetically between these two different structures and ground-state degeneracies will occur. Thus, similar to Li *et al.*, we simply excise this degeneracy by reducing the original 1081 structures to the 793 unique, nondegenerate solvation patterns.[51]

Geometrically, we have abstracted these two different energy models into an identical framework where the energy of sequence S in structure k, E_S^k, is simply a dot-product of the interaction vector $\boldsymbol{\gamma}_\mathbf{S}$ with the structure vector $\mathbf{D}^k$. $\boldsymbol{\gamma}_\mathbf{S}$ contains the interaction values for all pair–contact or solvation energies formed in a given sequence S. All sequence and amino-acid alphabet specific information such as the identity of amino-acids and amino-acid energetics is implicitly contained in and specified by $\boldsymbol{\gamma}_\mathbf{S}$. However, because we have restricted our attention to only maximally compact lattice conformations we can also abstract this framework to be amino-acid composition independent. Because all maximally compact lattice conformations have the same number of pair–contacts formed or buried residues, the following transformations on $\boldsymbol{\gamma}_\mathbf{S}$ will have no effect on the relative energy landscape

$$c_1\boldsymbol{\gamma}_\mathbf{S} \mapsto \boldsymbol{\gamma}_\mathbf{S}, \tag{5}$$

$$c_2\mathbf{1} + \boldsymbol{\gamma}_\mathbf{S} \mapsto \boldsymbol{\gamma}_\mathbf{S}, \tag{6}$$

where $\mathbf{1}$ is the identity vector. The scalar transformation in Eq. (5) is equivalent to shrinking or expanding the energy landscape by a multiplicative factor c_1, whereas the vector transformation in Eq. (6) is identical to boosting the energy landscape by a constant amount of energy c_2. The transformation described by Eq. (6) is the eigenvector spanning the nullspace of $\boldsymbol{\gamma}_\mathbf{S}$ across both energy models. For the pair–contact model the nullspace dimension is 1 and the eigenvector is $\mathbf{1}$, whereas the solvation model has a nullspace dimension of 2 and eigenvectors $\mathbf{1}_{\text{even}}=(0,1,0,...,1,0)$ and $\mathbf{1}_{\text{odd}}=(1,0,1,...,0,1)$. Based on these transformations, all $\boldsymbol{\gamma}_\mathbf{S}$ can be remapped with no adverse effect on the relative energy landscape and designability, so that

$$\boldsymbol{\gamma}_S \cdot \boldsymbol{\gamma}_S = 1, \tag{7}$$

$$\boldsymbol{\gamma}_\mathbf{S} \cdot \mathbf{1} = 0. \tag{8}$$

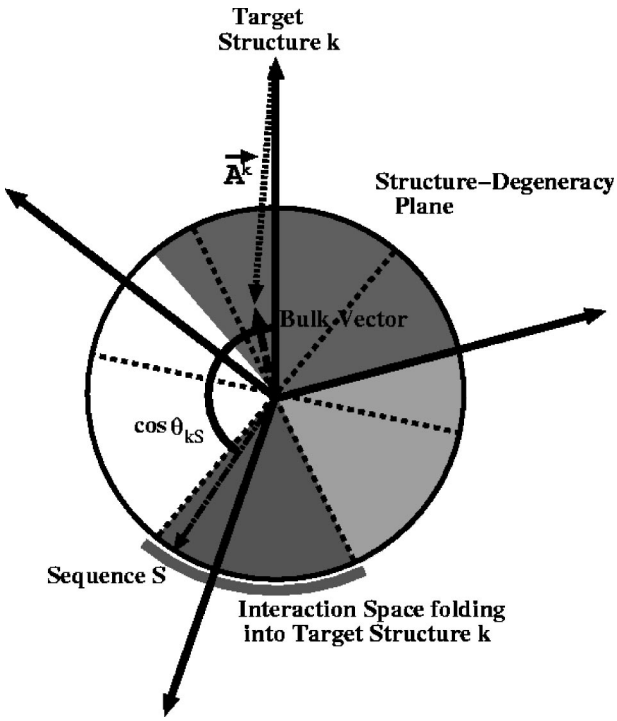

FIG. 2. A heuristic example demonstrating interaction space and explaining the geometric framework of sequence-structure energetics. The four structure vectors, $\boldsymbol{\mathcal{D}}$, are drawn as dark, solid arrows. The sample interaction vector, shown as a dot–dashed line near the bottom, folds into target structure k, as it has the farthest angular distance $\cos\theta_{kS}$ from k. The dotted lines are the structure-degeneracy planes between all structures and our target structure k. The area of total interaction space, farthest away from the native structure, folding into target structure k (drawn in the darkest shade) is bounded by a maximum of two of these structure-degeneracy planes. The other shaded arc-areas represent portions of interaction space folding into the other structures, respectively, farthest away from them. The small, dashed arrow is the bulk-vector $\langle\boldsymbol{\mathcal{D}}\rangle$ which is an average over all structure vectors. $\mathbf{A}^k$ spans between the target structure k and the bulk vector. Note that the largest areas of folding interaction space belong to those structures farthest away from the bulk vector and one another.

We have constrained an infinitely large interaction space $\boldsymbol{\gamma}$ to lie on the surface of a unit hypersphere whose ''center-of-mass'' is zero.[37] This universal vector framework, abstracted from a specific energy model and amino-acid composition, is inspired by and a synthesis of the work of Govindarajan and Goldstein and Tang and co-workers.[37,51] As a consequence, the energy for any given sequence S, independent of its specific composition, amino-acid alphabet size, or amino-acid interactions, lies between $-\mathcal{D}\leqslant E_S^k\leqslant\mathcal{D}$, where $\mathcal{D}$ is equal to the square root of the number of buried residues or pair–contacts for each compact 5×5 structure. As shown in Fig. 2

$$E_S^k = \|\boldsymbol{\gamma}_\mathbf{S}\|\|\boldsymbol{\mathcal{D}}^k\|\cos\theta_{kS} = \mathcal{D}\cos\theta_{kS}, \tag{9}$$

where $\cos\theta_{kS}$ is the projection angle between the normalized interaction vector $\boldsymbol{\gamma}_S$ and structure vector $\boldsymbol{\mathcal{D}}^k$. Assuming that all of the compact states are kinetically accessible, the native structure is the state of lowest free energy. Hence, in this geometrical framework, the native structure is that which has the lowest, unique E_S^k and $\mathcal{D}\cos\theta_{kS}$ or, equivalently, the structure $\boldsymbol{\mathcal{D}}^k$ which is farthest away from the interaction vector $\boldsymbol{\gamma}_S$. Similar to Govindarajan and Goldstein and Li *et al.*,

if one can further assume that the sequence space of *all* possible random, foldable protein sequences $\{\boldsymbol{\gamma}_S\}$ covers the entire unit-hypersphere interaction space isotropically, then the designability V_k of structure k breaks down into calculating the surface area of the interaction space unit-sphere of dimension d which folds into structure k: That surface area furthest from $\boldsymbol{\mathcal{D}}^k$ constrained by the appropriate structural-degeneracy planes. The structural-degeneracy planes are those hyper-dimensional planes of dimension $d-1$ in interaction space $\boldsymbol{\gamma}$ where two structures, $\boldsymbol{\mathcal{D}}^j$ and $\boldsymbol{\mathcal{D}}^k$, have identical energy. In terms of our designability calculation for a target structure k, one needs to determine the relevant $n-1$ structural-degeneracy planes which constrain a surface of interaction space $\boldsymbol{\gamma}$ farthest from $\boldsymbol{\mathcal{D}}^k$. However, note that those relevant structural-degeneracy planes will nearly always be determined by those structures closest in vector space to the target structure k. This observation forms the essence of the conclusions of Tang and co-workers; namely, highly designable structures are characterized by a low-density of surrounding structure vectors.[51] Calculating which structural-degeneracy planes are relevant for any given target structure has no simple analytical solution. Throughout this paper, we instead numerically calculate the designability of structures on the unit-hypersphere in interaction space via Monte Carlo integration. This is done by randomly drawing all d $\{\gamma_S\}$ interactions for the pair–contact or solvation model from a Gaussian distribution. In the theoretical protein folding literature, this Monte Carlo method is also known as the independent interaction model (IIM) or ''infinite''-amino-acid alphabet.[25,26,47]

FOLDABILITY, ENERGY GAP, AND DESIGNABILITY

Now that we have a universal model relating designability V_k to surface area in interaction space, the interesting question is how it relates to energetic measures, such as energy gap Δ_{10} and foldability $\mathcal{F}$. Previous work by Wolynes and co-workers derived a simple expression for the foldability $\mathcal{F}^k$ of sequence $\boldsymbol{\gamma}_S$ folding into structure k

$$\mathcal{F}^k(\boldsymbol{\gamma}_S)=\frac{\Delta_S^k}{\sigma_S}=\frac{\langle E_S\rangle-E_S^k}{\sqrt{\Sigma_l(\langle E_S\rangle-E_S^l)^2}}=\frac{\mathbf{A}_k\cdot\boldsymbol{\gamma}_S}{\sqrt{(\boldsymbol{\gamma}_S)^T\cdot\mathbf{B}\cdot\boldsymbol{\gamma}_S}}, \quad (10)$$

where $\langle E_S\rangle$ is the average energy of sequence S amongst all structures.[15,16,30] Note that the vector $\mathbf{A}_k$ and matrix $\mathbf{B}_{i,j}$ contain only structural-information

$$\mathbf{A}_k=\langle\boldsymbol{\mathcal{D}}\rangle-\boldsymbol{\mathcal{D}}^k \quad \text{and} \quad \mathbf{B}_{i,j}=\langle\mathcal{D}_i\mathcal{D}_j\rangle-\langle\mathcal{D}_i\rangle\langle\mathcal{D}_j\rangle, \quad (11)$$

where the averaging $\langle\cdots\rangle$ is done over all N structures. As shown in Fig. 2, $\mathbf{A}_k$ is a vector spanning between the target structure k and the ''bulk'' structural vector $\langle\boldsymbol{\mathcal{D}}\rangle$. The simple interpretation of the bulk vector is that it counts how frequent (common) or infrequent (rare) certain pair–contacts or solvated residues are amongst the ensemble of possible structures. Clearly, not all $\mathbf{A}_k$ are of equal length as some structures k are farther away from the bulk than others. In the limit of large N, $\|\langle\boldsymbol{\mathcal{D}}\rangle\|$ is $\sim\mathcal{D}^2/\sqrt{d}$. Because all $\mathbf{D}^k$ are of equal magnitude, those structures which are furthest away from the bulk or, equivalently, those structures which contain the maximum amount of rare pair–contacts or buried residues are those with larger $\mathbf{A}_k$. Since the bulk energy gap is given by

$$\Delta_S^k=\mathbf{A}_k\cdot\boldsymbol{\gamma}_S=\|\mathbf{A}_k\|\cos\theta_{A_kS}, \quad (12)$$

it is immediately obvious that those structures with a maximum of rare pair–contacts or solvated residues are also the most optimizable, or, equivalently $\Delta_{\text{opt}}^k=\|\mathbf{A}_k\|$. However, Govindarajan and Goldstein were interested in foldability $\mathcal{F}^k(\boldsymbol{\gamma}_S)$, whose denominator is a function of $\mathbf{B}$ and γ_S. $\mathbf{B}_{i,j}$ is a matrix representing the correlations between pair–contacts or buried residues amongst the ensemble of all structures. Namely, $\mathbf{B}_{i,j}$ is positive or negative if the pair–contact set or buried residue i occurs in conjunction with j more or less often than what is expected at random given the pair–contact or buried residue frequencies found in the bulk vector. From Eq. (10) and the appropriate $\mathbf{A}_k$ and $\mathbf{B}_{i,j}$ of the energy model, the optimal foldability $\mathcal{F}_{\text{opt}}^k$ for a target structure k can be obtained in closed form[15,16,30,32]

$$\boldsymbol{\nabla}\mathcal{F}^k(\boldsymbol{\gamma}_{\text{opt}})=0\rightarrow\boldsymbol{\gamma}_{\text{opt}}=\mathbf{B}^{-1}\cdot\mathbf{A}_k$$

$$\rightarrow\mathcal{F}_{\text{opt}}^k=\sqrt{(\mathbf{A}_k)^T\cdot\mathbf{B}^{-1}\cdot\mathbf{A}_k}. \quad (13)$$

Both measures $\mathcal{F}$ and Δ_{10} are simply related to designability in that any energy gap, whether Δ^k of foldability or Shakhnovich's Δ_{10}, is but a projection of $\boldsymbol{\gamma}_S$ onto some spanning vector $\boldsymbol{\mathcal{D}}\star-\boldsymbol{\mathcal{D}}^k$, which always originates from the global-energy minimum native state k. For Δ^k, this spanning vector is exactly $\mathbf{A}_k$. Hence, those structures farthest away from the bulk-vector $\langle\boldsymbol{\mathcal{D}}\rangle$ will tend to have sequences folding into them with larger Δ^k and, neglecting $\mathbf{B}$ terms, the foldability $\mathcal{F}$. For Δ_{10}, the spanning vector is between the native state and the next farthest structure l from $\boldsymbol{\gamma}_S$. Unfortunately for analytical purposes, $\boldsymbol{\mathcal{D}}^l$ is sensitive to and dependent on both $\boldsymbol{\mathcal{D}}^k$ and $\boldsymbol{\gamma}_S$. Notice, however, that $\boldsymbol{\mathcal{D}}^l-\boldsymbol{\mathcal{D}}^k$ is but the vector defining the structural-degeneracy plane of both the native state k and l. Thus, coming full circle to previous arguments regarding designability V_k, those structures with a minimal density of surrounding structure vectors will also be those with sequences folding into them with larger possible Δ_{10}. The connection relating both $\mathcal{F}$ and Δ_{10} to designability V_k is based on the aforementioned observation that those structures farthest away from the bulk-vector $\langle\boldsymbol{\mathcal{D}}\rangle$ also tend to have the smallest density of surrounding structure vectors on the Euclidean hyper-cube. Thus, given that these energetic measures $\mathcal{F}$ and Δ_{10} necessarily reflect the underlying distribution and density of $\boldsymbol{\mathcal{D}}^k$, is it really surprising that they be correlated to designability?

As a synthesis of this relation between low structural density, increased distance from the bulk, and higher designability V_k, we propose a novel structural measure Z_d^k. The main impetus behind this measure was to incorporate low structural density as a critical feature for large structural designability. Our measure is based on the Euclidean distance between any two structural vectors l and k, $d_{kl}=\|\boldsymbol{\mathcal{D}}^l-\boldsymbol{\mathcal{D}}^k\|$ and $0\leqslant d_{kl}\leqslant 2\mathcal{D}$. Given a target structure k and the density of its distance from other structures $\rho(d_{kj})$

TABLE I. Statistical parameters describing the percentage ground-state degeneracy and the Pearson correlation coefficients between various measures $\mathcal{F}_{\text{opt}}^k$, Z_d^k, $\langle\mathcal{F}\rangle$, $\langle\Delta_{10}\rangle$ and structure designability V_k across different energy models and amino-acid alphabets. These correlation coefficients are numerical supplements to Figs. 3–10.

Alphabet and energy model	% Degeneracy	$\mathcal{F}_{\text{opt}}^k$ vs V_k	Z_d^k vs V_k	$\langle\mathcal{F}\rangle$ vs V_k	$\langle\Delta_{10}\rangle$ vs V_k
HP two-letter pair–contact	81.58%	0.583	0.501	0.036	0.740
Li two-letter pair–contact	63.09%	0.414	0.345	0.442	0.783
π two-letter pair–contact	61.65%	0.489	0.416	0.287	0.661
hXYX four-letter pair–contact	41.63%	0.669	0.630	0.279	0.928
MJ 20-letter pair–contact	4.39%	0.829	0.885	0.924	0.943
Monte Carlo IIM pair–contact	0.01%	0.828	0.907	0.942	0.926
HP two-letter solvation	84.72%	0.505	0.704	0.272	0.585
FVSQ four-letter solvation	45.40%	0.538	0.795	0.591	0.703
H_2O–Octanol 20-letter solvation	8.58%	0.522	0.843	0.786	0.879
Monte Carlo IIM solvation	1.64%	0.603	0.868	0.853	0.915

$$Z_d^k=\frac{\Delta_{d_{kl}}}{\sigma_{d_{kl}}}=\frac{\langle d_k\rangle-d_{kk}}{\sqrt{\Sigma_l(\langle d_k\rangle-d_{kl})^2}}=\frac{\langle d_k\rangle}{\sqrt{\Sigma_j(\langle d_k\rangle-d_{kl})^2}}, \quad (14)$$

where $\langle d_k\rangle=(1/N)\Sigma_l d_{kl}$. Maximization of the target structure k distance from the bulk is handled by the numerator, while the minimization of density of close, competing structures is implicitly taken into account by the denominator. Thus, similar in nature to optimal foldability $\mathcal{F}_{\text{opt}}$, we posit that highly designable structures k will have larger Z_d^k. The strong appeal of Z_d^k is that it is exceedingly easy to calculate, unlike $\mathcal{F}_{\text{opt}}^k$ which involves carefully inverting a high-dimensional matrix **B**. Furthermore, our results demonstrate that Z_d^k is generally better correlated to designability V_k than $\mathcal{F}_{\text{opt}}^k$ across both energy models. However, unlike foldability and its relation to T_f/T_g, Z_d^k is not obviously interpretable in any energetic or thermodynamic sense.

METHODS AND RESULTS

For each amino-acid alphabet and energy model, we ensured that the native state was the nondegenerate ground state amongst all possible compact structures. The 5×5 2D pair–contact model has a total of 1081 unique structures and there are 793 unique, nondegenerate structures for the 5×5 solvation model. Within the pair–contact model, we used six amino-acid alphabets: The HP two-letter, Li two-letter, π two-letter, hHYX four-letter, MJ 20-letter, and the Monte Carlo (IIM) ''infinite''-letter amino-acid alphabet. The interaction details for these amino-acid alphabets are described elsewhere.[50] For the solvation model, we used 4 amino-acid alphabets: An HP two-letter, FVSQ four-letter, H_2O–Octanol 20-letter, and a Monte Carlo (IIM) infinite-letter amino-acid alphabet. The HP two-letter is identical to that used by Li and co-workers and analyzed by Ejtehadi and co-workers, where a buried hydrophobe is energetically favored $[\gamma(H)=-1,\gamma(P)=0]$.[51–53] Both FVSQ four-letter and 20-letter amino-acid alphabets were taken from experimentally measured amino-acid water-to-octanol ΔG's as given by Roseman.[60] FVSQ, a rough spectrum along the lines of bulkiness and hydrophobicity, represents the four naturally occurring amino-acids. Both Monte Carlo alphabets had each interaction, whether pair–contact or solvation, independently drawn from a Gaussian distribution. As previous mentioned, Monte Carlo sampling represents a numerical calculation of the exact structure designabilities V_k in their respective energy models in the limit of isotropic distribution of sequences in interaction space.

For all two-letter amino-acid alphabets, we exhaustively screened every possible sequence ($2^{25}\sim33$ million), whereas $\sim$20 million sequences were selected at random for the remaining higher amino-acid alphabets. Conformations were considered degenerate if their energies differed by less than 10^{-4}. In all cases, our sampling was large enough to suppress statistical fluctuation. Naturally, given the larger diversity of interactions, random sequences constructed from larger amino-acid alphabets had less ground-state degeneracies than their smaller counterparts (See Table I). For each sequence folding into a ground-state structure, we calculated the foldability $\mathcal{F}$ and energy gap Δ_{10} of that sequence. Normally, all $\mathcal{F}$ are dimensionless and amino-acid composition independent, but for explicit comparison to previous designability results we left Δ_{10} unnormalized and sequence-dependent.[29] $\mathcal{F}_{\text{opt}}^k$ and Z_d^k [as defined in Eqs. (13) and (14)] were also calculated for each structure k within their respective energy model. For aesthetic reasons, we normalized the structure designabilities within a given amino-acid alphabet with the following constraint:

$$\sum_k V_k=10\,000. \quad (15)$$

We begin by looking at the pair–contact model results. Validating the basic premise of Govindarajan and Goldstein, Fig. 3 is a plot of $\mathcal{F}_{\text{opt}}^k$ versus designability V_k. The striking feature is the positive correlation across all amino-acid alphabets, although significantly better for higher-letter amino-acid alphabets (all statistical details are in Table I). Figure 4 is a plot of Z_d^k versus designability V_k. As expected in its construction, Z_d^k exhibits a better correlation to designability than $\mathcal{F}_{\text{opt}}^k$; again, the correlation is significantly better pronounced for higher-letter alphabets. However, even for higher-letter alphabet, the correlation between $\mathcal{F}_{\text{opt}}^k$ or Z_d^k and V_k is not perfect. Realistically, one could predict *a priori* with confidence which lattice structures belong to which tier: Lowly designable, moderately designable, or highly design-

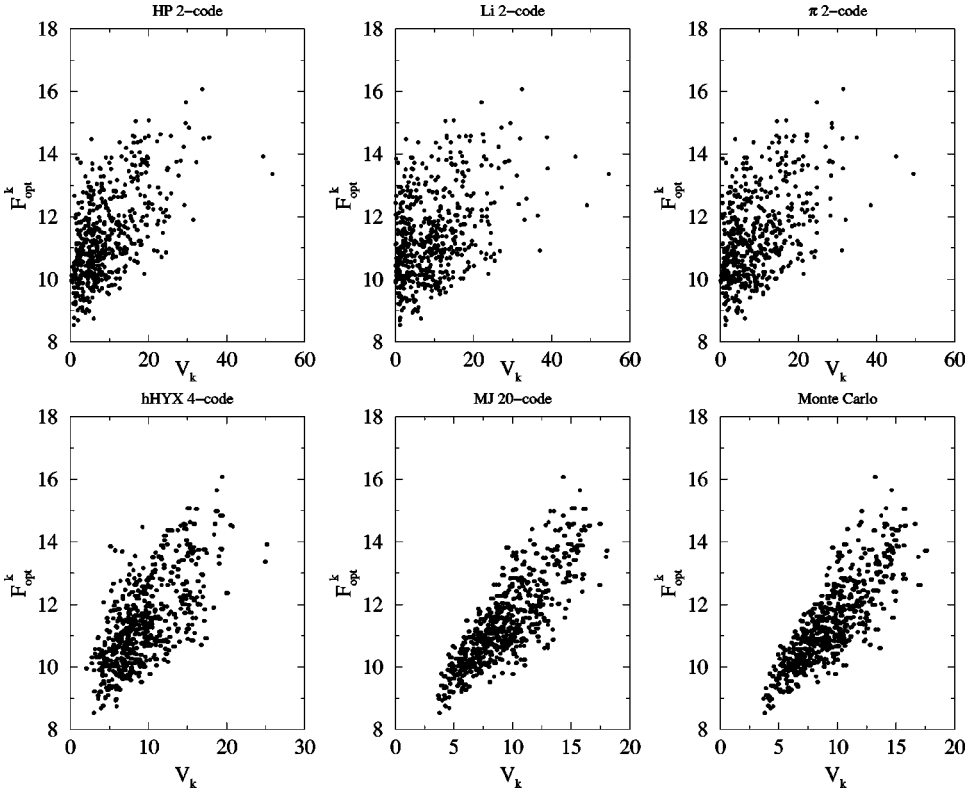

FIG. 3. A plot of $\mathcal{F}^k_{\rm opt}$ vs designability V_k, calculated for various amino-acid alphabets in a pair–contact model. The amino-acid alphabet is listed above each respective plot.

able. This coarse resolution is an unavoidable trade-off for using simple, scalar measures and avoiding the explicit enumeration of all possible foldable sequences.

Concerning thermodynamically relevant measures, $\langle\mathcal{F}\rangle$ and $\langle\Delta_{10}\rangle$, we begin by earmarking those structures which are nondegenerate within Ejtehadi and co-workers solvation-like, linear interaction model ($\gamma_M=0$) for explicit comparison to their results. There are only 661 of 1081 structures with unique interaction patterns, which are drawn as large triangles in Figs. 5 and 6. For two-letter amino-acid alphabets, there is a striking division between highly designable triangles and the remaining lowly designable structures. The

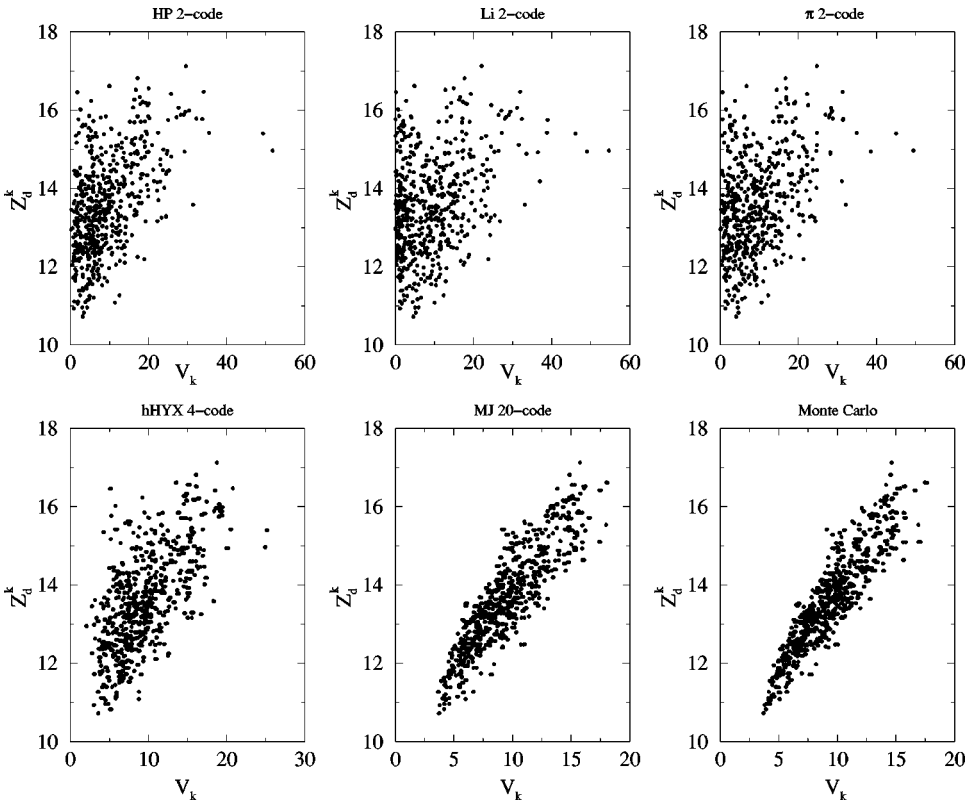

FIG. 4. A plot of Z^k_d vs designability V_k, calculated for various amino-acid alphabets in a pair–contact model. The amino-acid alphabet is listed above each respective plot.

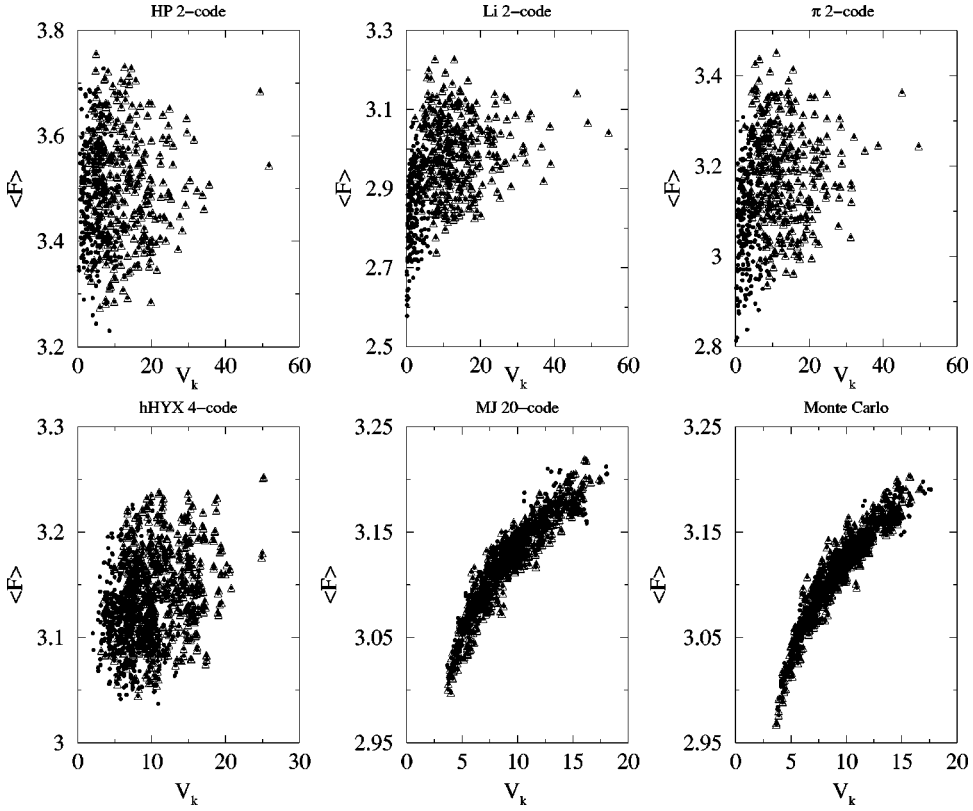

FIG. 5. A plot of $\langle \mathcal{F} \rangle$ vs designability V_k, calculated for various amino-acid alphabets in a pair–contact model. The large triangles surround those 661 structures which are nondegenerate in the $\gamma_M = 0$ linear interaction model of Ejtehadi and co-workers.

significance of this segregation disappears for higher-letter amino-acid alphabets, as there is large-scale rearrangement in the relative designability of structures. The correlation between $\langle \mathcal{F} \rangle$ and V_k exhibits dramatic sensitivity to amino-acid alphabet size. Figure 6, a plot of $\langle \Delta_{10} \rangle$ versus V_k, is consistent with results of Ejtehadi and co-workers and Tang and co-workers;[33,52,53] namely, two-letter amino-acid alphabets with a small mixing parameter γ_M, such as the Li two-letter code, have a discontinuous jump in $\langle \Delta_{10} \rangle$ between highly designable and lowly designable structures. However, when $\gamma_M > \gamma_{\text{crit}} \sim 1$, as for the HP two-letter and π two-letter code, this discontinuity mostly disappears. Thus, in light of previous results, $\gamma_M > \gamma_{\text{crit}}$ does not appear to affect the structure designability of two-letter amino-acid alphabets, but rather only the statistics of $\mathcal{F}$ and $\langle \Delta_{10} \rangle$ with regard to V_k.[52,53,50] The discontinuous jump in $\langle \Delta_{10} \rangle$ as observed by Li *et al.* is clearly an artifact of having a two-letter amino-acid alphabet that is but a small perturbation of the solvationlike linear interaction model. However, unlike $\langle \mathcal{F} \rangle$, all amino-acid alphabets maintain a significantly positive correlation between $\langle \Delta_{10} \rangle$ and V_k, although much better for higher-letter alphabets. All in all, in conjunction with our previous paper,[50] these results indicate that two-letter amino-acid alphabets have irrefutable differences, not only with which structures are highly designable (see Fig. 11), but also in $\mathcal{F}^k_{\text{opt}}$, Z^k_d, $\langle \mathcal{F} \rangle$, and $\langle \Delta_{10} \rangle$ versus V_k, when compared to higher-letter amino-acid alphabets for pair-contact models. Of notable mention, the MJ 20-letter and Monte Carlo alphabet have nearly identical results and both reflect the exact pair–contact model unlike the smaller-letter codes. In short, two-letter amino-acid alphabets are plagued by artifacts that are in contrast to the exact pair–contact model.

The solvation model data, shown in Figs. 7–10, exhibit identical results to the pair–contact model for V_k versus $\mathcal{F}^k_{\text{opt}}$, Z^k_d, $\langle \mathcal{F} \rangle$, and $\langle \Delta_{10} \rangle$ as a function of amino-acid alphabet size. However, there are a few notable differences worth exploring. As shown in Fig. 11 and 12, unlike the pair–contact model, the designability V_k of specific solvation structures across these different amino-acid alphabets *does not* significantly change! Namely, those structures which are highly designable for the HP two-letter solvation model are also highly designable for the Monte Carlo alphabet. Additionally, unlike the pair–contact model, the correlation between Z^k_d and V_k in Fig. 8 is remarkably better than that of $\mathcal{F}^k_{\text{opt}}$. Paralleling this observation, the breakdown between Z^k_d and $\mathcal{F}^k_{\text{opt}}$ is shown in Fig. 13, which is a plot of these two measures against one another in the pair–contact and solvation model. Clearly, $\mathcal{F}^k_{\text{opt}}$ and Z^k_d are correlated to one another, but remarkably more so for the pair–contact model. This better correlation may be a consequence of the sparseness of common interactions (even ''rarer'') of the pair–contact model (16 pair–contacts/132 possible) compared to the solvation model (9 buried/25 possible). This phenomenon is certainly responsible for the larger range of $\mathcal{F}^k_{\text{opt}}$ and Z^k_d values for the pair–contact model compared to the solvation model.

Given our suspicion that two-letter amino-acid alphabets in a pair–contact model could be solvation models in disguise, how do the relative designabilities compare *across* energy models? Figure 14 is a scatter plot of the Monte Carlo solvation model designability $V_k^{\text{MC solution}}$ versus the structure designabilities for the six different amino-acid alphabets in a pair–contact model. Even the hHYX four-letter code design-

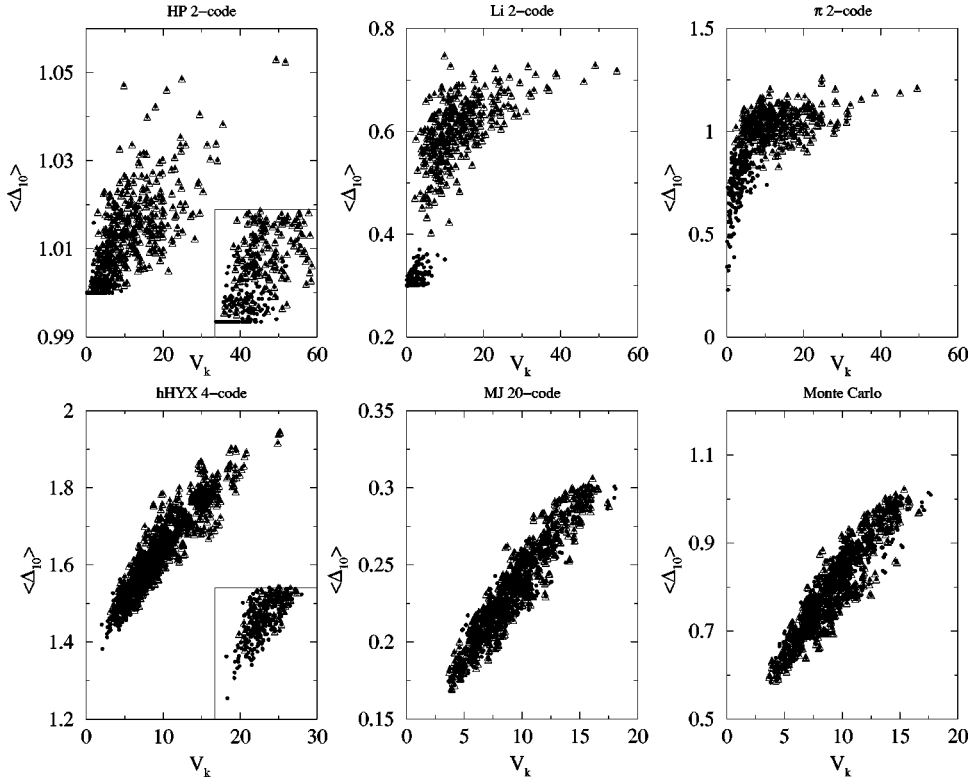

FIG. 6. A plot of $\langle \Delta_{10} \rangle$ vs designability V_k, calculated for various amino-acid alphabets in a pair–contact model. The large triangles are those 661 structures which are nondegenerate in the $\gamma_M = 0$ linear interaction model of Ejtehadi and co-workers.

ability exhibits strong correlation to the solvation model, particularly those highly designable structures. Thus, in retrospect, the HP two-letter designability results of Li and coworkers concerning proteinlike symmetries of highly designable structures are a consequence of a hidden solvation model, rather than reflecting properties endemic to a pair–contact model.

Concerning Kussell and Shakhnovich's conclusions

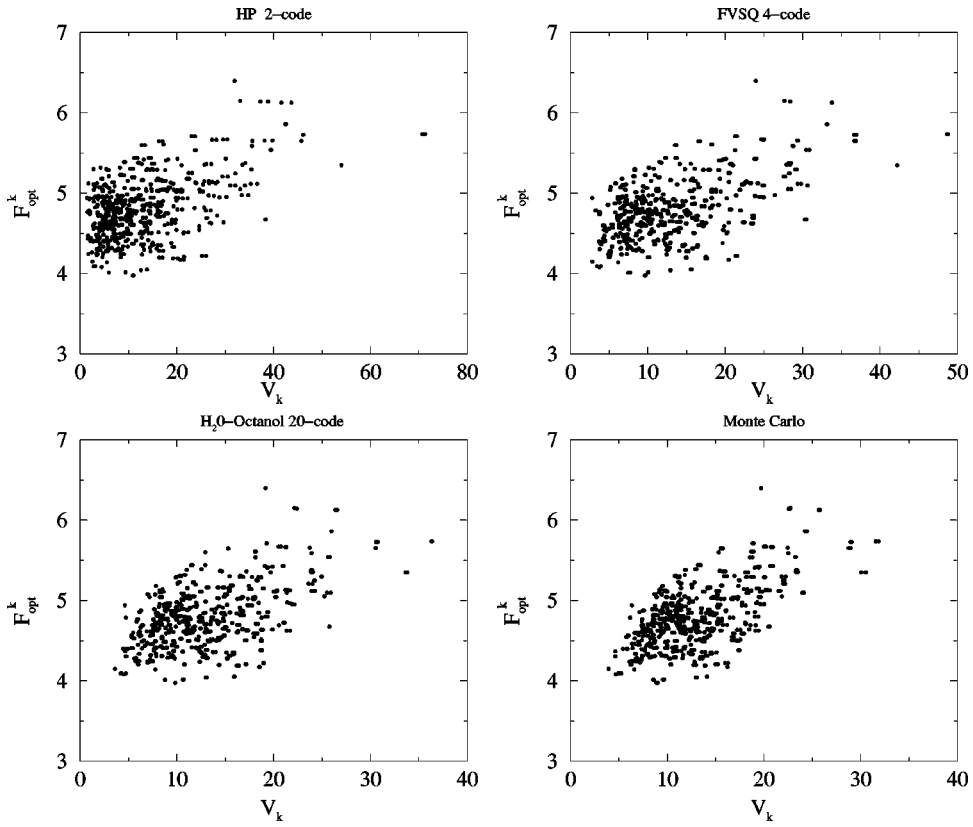

FIG. 7. A plot of $\mathcal{F}^k_{\rm opt}$ vs designability V_k, calculated for various amino-acid alphabets in a solvation model.

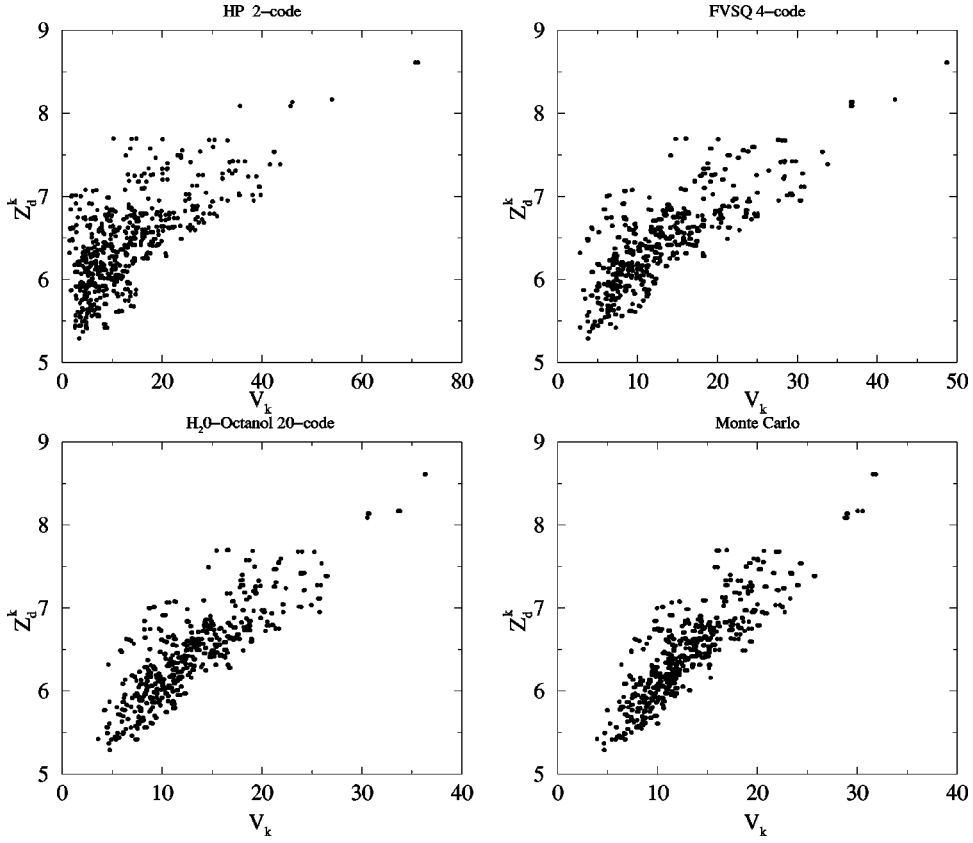

FIG. 8. A plot of Z_d^k vs designability V_k, calculated for various amino-acid alphabets in a solvation model.

about 2D structural features pertinent to designability for a pair–contact model, we also calculated the number of interaction loops and strands for every pair–contact structure and divided them into six classes: loops, one-length, two-length, three-length, four-length, and five+-length strands. Within a class, we averaged over the designability $\langle V_k \rangle$ for those structures containing a certain number of loops, one-length strands, etc. Kussell and Shakhnovich postulated that for 2D

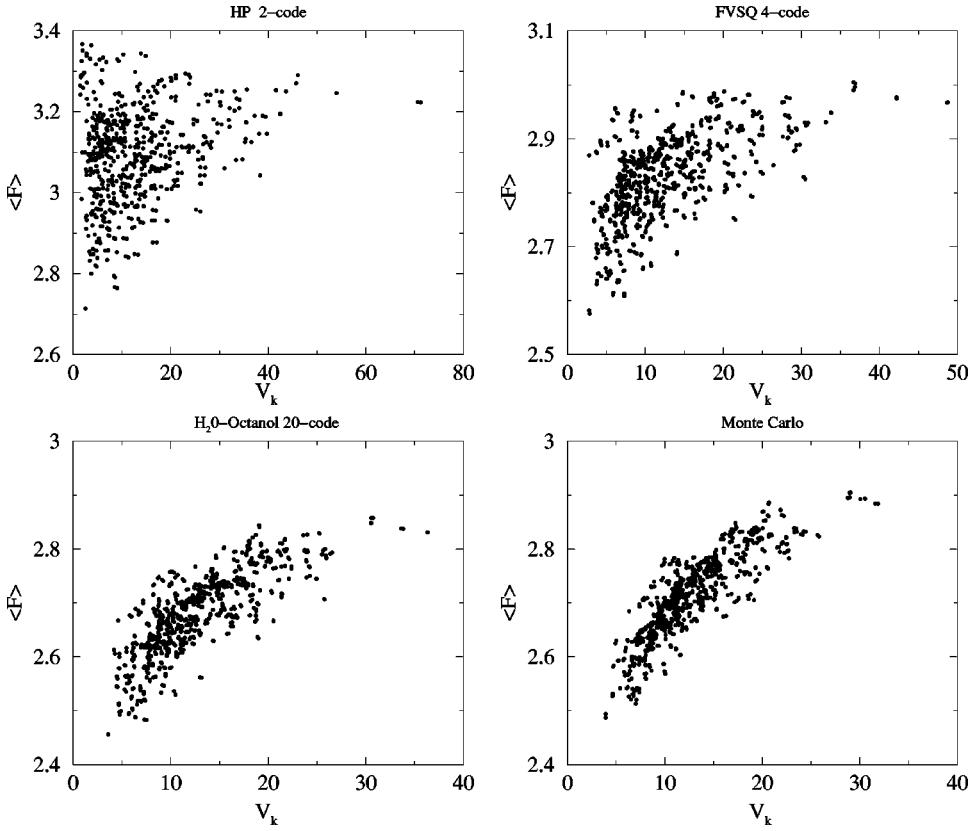

FIG. 9. A plot of $\langle \mathcal{F} \rangle$ vs designability V_k, calculated for various amino-acid alphabets in a solvation model.

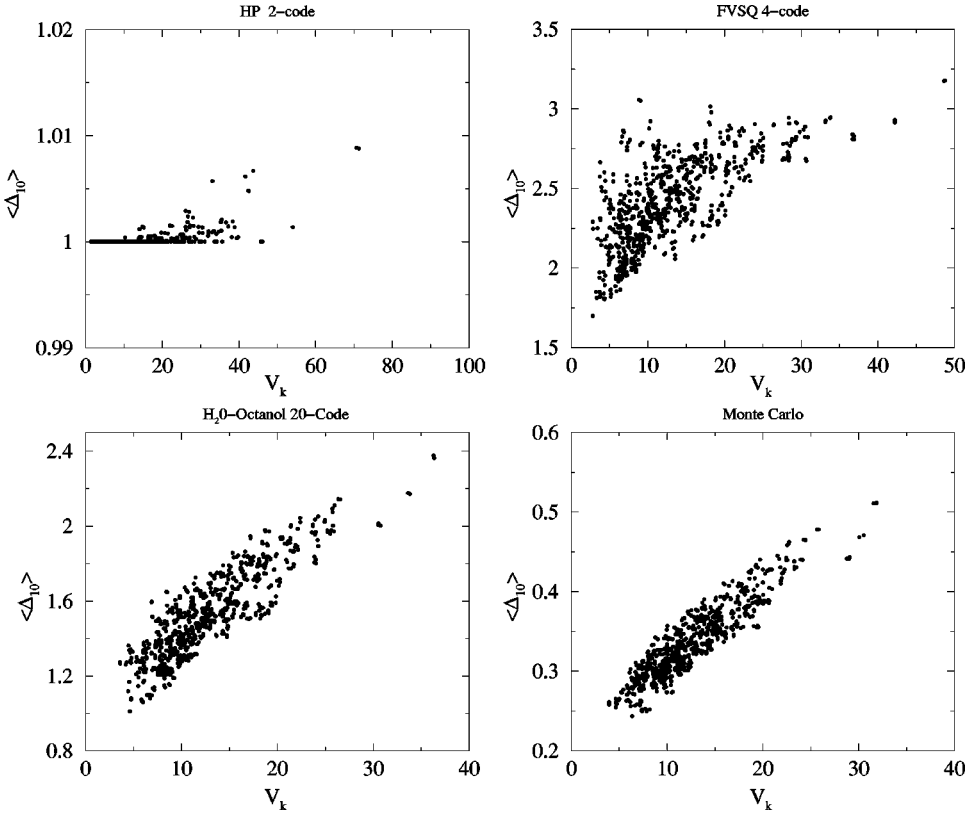

FIG. 10. A plot of $\langle \Delta_{10} \rangle$ vs designability V_k, calculated for various amino-acid alphabets in a solvation model.

lattice proteins constructed with a two-letter amino-acid alphabet, highly designable structures were expected to have (1) no loops, (2) a maximum number of two-length strands, and (3) a minimum number of larger-length strands. Our results are shown in Fig. 15. Of the three conditions set forth by Kussell and Shakhnovich for two-letter alphabets, only loops are in agreement. Strangely, the number of two-length strands for two-letter amino-acid alphabets gives vague and contradictory correlations to designability. In complete disagreement with Kussell and Shakhnovich, having a larger

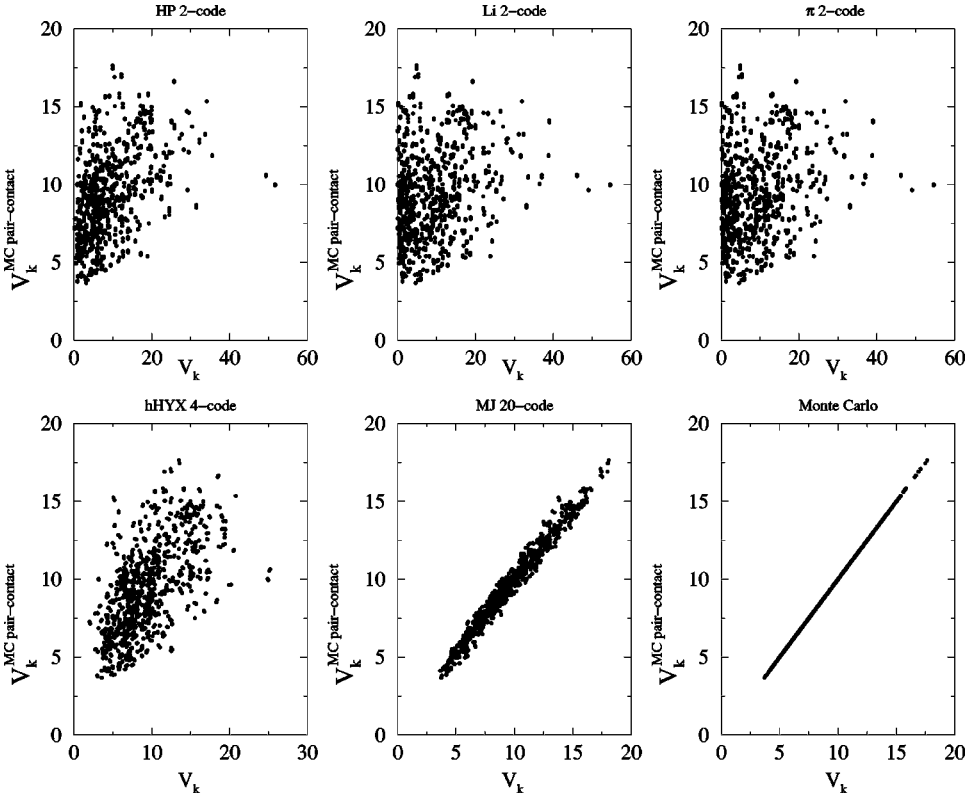

FIG. 11. A plot of the exact designability $V_k^{\text{MC pair–contact}}$ vs various amino-acid alphabets, listed above each respective plot, in a pair–contact model.

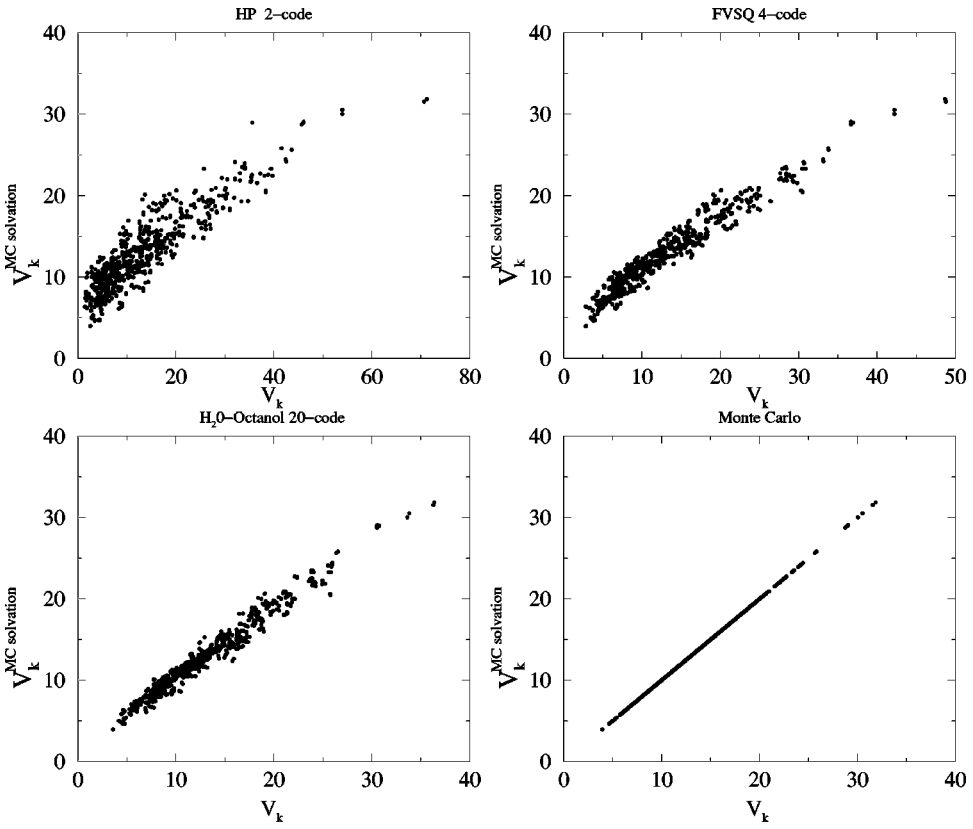

FIG. 12. A plot of the exact designability $V_k^{\text{MC solvation}}$ vs various amino-acid alphabets, listed above each respective plot, in a solvation model.

number of three-length and four-length strands is synonymous with higher designability for these two-letter alphabets. A possible source of two-letter alphabet deviation of this analytical theory from our lattice protein simulation involves the use of the REM assumption by Kussell and Shakhnovich to translate the energy spectrum $n^k(E)$, the number of sequences having energy E for structure k, into designability. Their analytical results are contingent on the validity of the REM. Yet, it is important to note that the REM is questionable for two dimensional systems: That is, both for a replica derivation of random heteropolymers in 2D[12,58] and for a nonreplica derivation from loop entropy arguments for designed, proteinlike sequences.[47] In addition, it has also been conjectured many times and numerically shown by Pande and co-workers that the REM works far better for higher-letter alphabets than for two-letter alphabets.[61] Thus, it should not be too surprising that the Kussell and Shakhnovich theory in 2D with a two-letter code does not agree very well with our lattice protein results. However, if Kussell and Shakhnovich's analytical model is more valid for higher-letter alphabets, Fig. 15 hints that the following features for 2D lattice proteins should be strongly correlated with larger designability for higher-letter alphabets: (1) No loops, (2) a minimum of one-length strands, and (3) a maximum of two-length and three-length strands. Granted, four-length strands also give positive correlations to higher designability for larger-letter alphabets, but their signature is much weaker in comparison.

CONCLUSIONS

Our results lend strong credence to the fact that we are starting to understand the principles of what makes a protein structure designable. Based on the work by Govindarajan

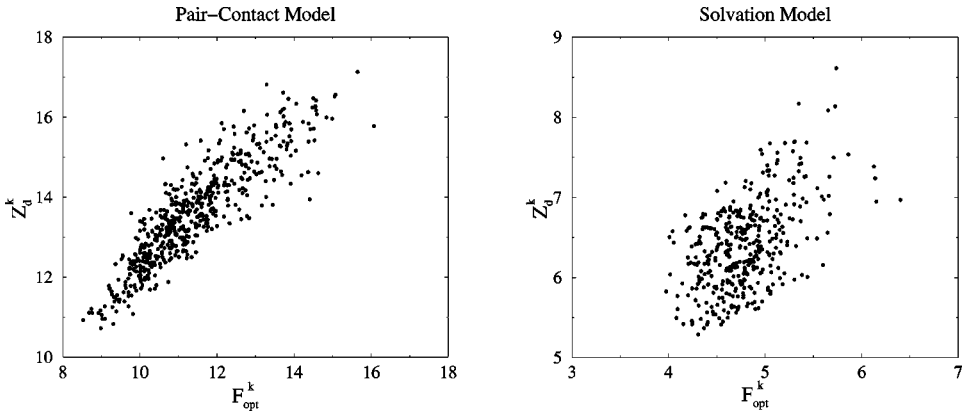

FIG. 13. A plot of $\mathcal{F}_{\text{opt}}^k$ vs Z_d^k across both energy models: Pair–contact and solvation.

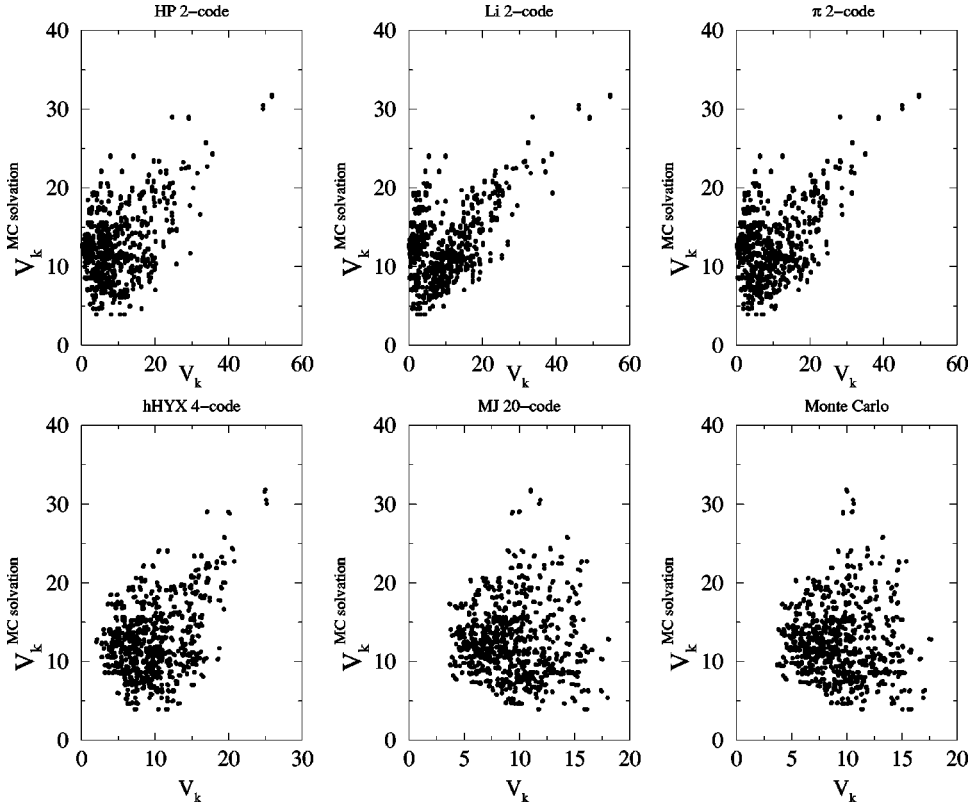

FIG. 14. A plot of the Monte Carlo designability $V_k^{\text{MC solvation}}$ for the solvation model vs that across all amino-acid alphabets (listed above each respective plot) for the pair–contact model. These data substantiate the claim that two-letter amino-acid alphabets in the pair–contact model are solvation models in disguise.

and Goldstein with pair–contact models and Li *et al.* with solvation models, those structures farthest away from the bulk, which are also those with the smallest density of surrounding structure vectors, are highly designable. Structurally-calculable measures $\mathcal{F}_{\text{opt}}^k$ and Z_d^k in Figs. 3 and 4 and 7 and 8 demonstrate this principle across pair–contact and solvation models. Naturally, *which* structures are highly designable depends on the particular energy model used. For the pair-contact model, highly designable structures have an abundance of rare, long-range pair–contacts, whereas for the

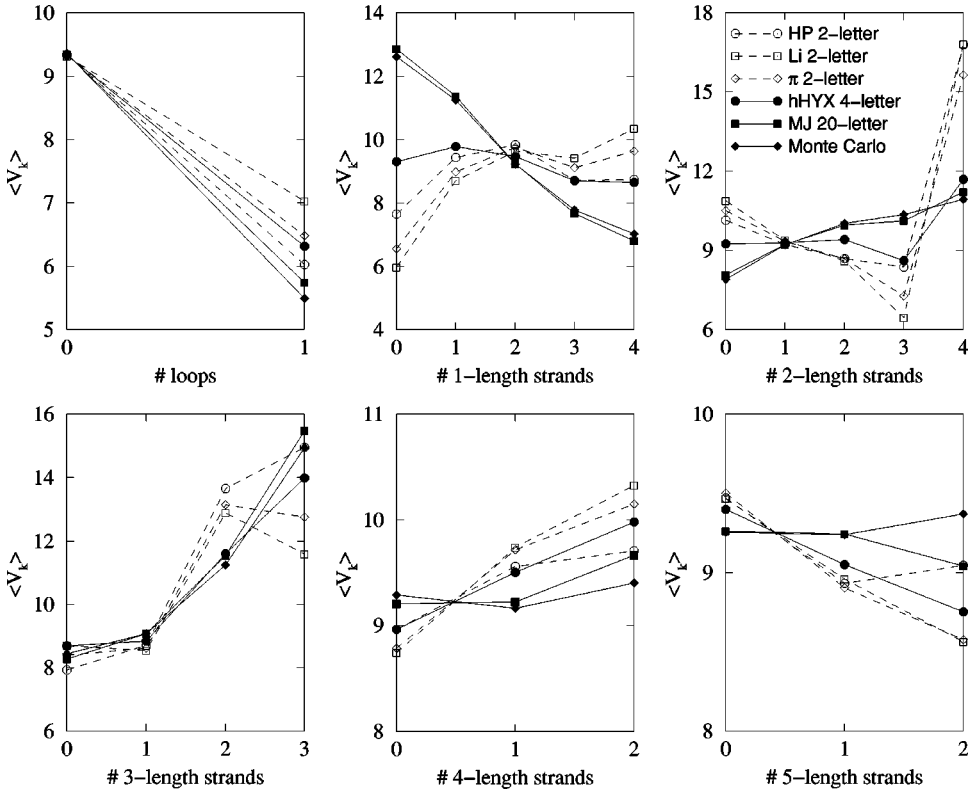

FIG. 15. A decomposition of pair–contact structures according to number of loops, one-length, two-length, three-length, four-length, and five+-length interaction strands plotted vs the average designability $\langle V_k \rangle$ over structures containing these particular features. The data from the various alphabets are drawn with different curves and symbols, as described by the legend in the two-strand plot.

solvation model highly designable structure have many rare solvated residues found in those structures with proteinlike symmetries. In addition, consistent with the foldability designability model, Figs. 5 and 6 and 9 and 10 demonstrate that both $\langle \mathcal{F} \rangle$ and $\langle \Delta_{10} \rangle$ are sharply increasing functions of designability V_k across both energy models. Hence, increased thermodynamic stability and faster folding is a necessary, universal correlate with larger designability. Thus, in light of the hasty claim that the ''designability principle'' is an alternative to the foldability model,[62] our results indicate that they are inherently related across different energy models. This correlation has ramifications for protein evolution of thermodynamic or folding properties on neutral networks, as recently applied to directed evolution.[63]

Consistent with our previous report, the correlation between V_k and $\mathcal{F}^k_{\mathrm{opt}}$, Z^k_d, $\langle \mathcal{F} \rangle$, $\langle \Delta_{10} \rangle$ consistently breaks down for smaller-letter amino-acid alphabets in both energy models. Universally, smaller-letter codes contain artifacts that lead to deviant behavior from that of the exact energy model, as calculated by Monte Carlo sampling. The real controversy and open problem in protein designability lies in understanding why two-letter alphabet results break down for a given energy model and explaining their differences from higher-letter alphabets. Of specific interest, it was shown that the designability results of two-letter amino-acid alphabets for the pair-contact model are consistent with their being solvation models in disguise. This was delineated by Ejtehadi and co-workers and explicitly shown in Fig. 14. This conveniently explains the puzzle of why highly designable HP structures had proteinlike symmetries in spite their being pair–contact models. Given that most naturally occuring proteins have amazing fold symmetries, it is tempting to speculate that this could primarily be due to dominant solvation forces. In particular, it would be interesting to survey protein folds and their prevalence in the database and crosscheck against their Z^k_d within a solvation model.

ACKNOWLEDGMENTS

We would like to thank Darin Taverna, Manuel Blickle, Eugene Shakhnovich, and Reza Ejtehadi for helpful discussions and Todd Raeker for computational support. Financial assistance was provided by NIH Grant Nos. LM05770 and GM08270, and NSF shared-equipment Grant No. BIR9512955.

[1] M. Levitt and C. Chothia, Nature (London) **261**, 552 (1976).
[2] C. A. Orengo, D. T. Jones, and J. M. Thornton, Nature (London) **372**, 631 (1994).
[3] A. G. Murzin, S. E. Brenner, T. J. P. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[4] A. V. Finkelstein and O. B. Ptitsyn, Prog. Biophys. Mol. Biol. **50**, 171 (1987).
[5] A. V. Finkelstein and B. Reva, Nature (London) **351**, 497 (1991).
[6] A. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, FEBS Lett. **325**, 23 (1993).
[7] A. V. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, Subcell Biochem. **24**, 1 (1995).
[8] A. V. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, Proteins **23**, 142 (1995).
[9] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[10] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[11] E. I. Shakhnovich and A. M. Gutin, J. Phys. A **22**, 1647 (1989).
[12] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[13] E. I. Shakhnovich and A. V. Finkelstein, Europhys. Lett. **9**, 569 (1989).
[14] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 171 (1990).
[15] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).
[16] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 9029 (1992).
[17] J. U. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164 (1991).
[18] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).
[19] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).
[20] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[21] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA **92**, 1282 (1995).
[22] H. Nymeyer, A. Garcia, and J. Onuchic, Proc. Natl. Acad. Sci. USA **95**, 5921 (1998).
[23] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[24] M. R. Betancourt and J. N. Onuchic, J. Chem. Phys. **103**, 773 (1995).
[25] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[26] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[27] A. Šali, E. I. Shakhnovich, and M. J. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[28] A. Šali, E. I. Shakhnovich, and M. J. Karplus, Nature (London) **369**, 248 (1994).
[29] N. E. G. Buchler and R. A. Goldstein, J. Chem. Phys. **111**, 6599 (1999).
[30] S. Govindarajan and R. A. Goldstein, Biopolymers **36**, 43 (1995).
[31] S. Govindarajan and R. A. Goldstein, Proc. Natl. Acad. Sci. USA **93**, 3341 (1996).
[32] S. Govindarajan and R. A. Goldstein, Proteins **22**, 413 (1995).
[33] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[34] J. D. Bryngelson, J. Chem. Phys. **100**, 6038 (1994).
[35] V. S. Pande, A. Y. Grosberg, and T. Tanaka, J. Chem. Phys. **103**, 9482 (1995).
[36] S. Govindarajan and R. A. Goldstein, Biopolymers **42**, 427 (1997).
[37] S. Govindarajan and R. A. Goldstein, Proteins **29**, 461 (1997).
[38] E. Bornberg-Bauer, Biophys. J. **73**, 2393 (1997).
[39] M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **78**, 3967 (1997).
[40] M. Vendruscolo, Physica A **249**, 576 (1998).
[41] E. D. Nelson and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **95**, 10682 (1998).
[42] G. Trinquier and Y. H. Sanejouand, Phys. Rev. E **59**, 942 (1999).
[43] R. A. Broglia *et al.*, Phys. Rev. Lett. **82**, 4727 (1999).
[44] R. Mélin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[45] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. USA **96**, 10689 (1999).
[46] D. M. Taverna and R. A. Goldstein, Biopolymers **53**, 1 (2000).
[47] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Biophys. J. **73**, 3192 (1997).
[48] A. M. Gutin and E. I. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[49] E. I. Shakhnovich, Folding Des. **3**, R45 (1998).
[50] N. E. G. Buchler and R. A. Goldstein, Proteins **34**, 113 (1999).
[51] H. Li, C. Tang, and N. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998).
[52] M. R. Ejtehadi *et al.*, Phys. Rev. E **57**, 3298 (1998).
[53] M. R. Ejtehadi *et al.*, J. Phys. A **31**, 6141 (1998).
[54] M. R. Ejtehadi, N. Hamedani, and V. Shahrezaei, Phys. Rev. Lett. **82**, 4723 (1999).
[55] V. Shahrezaei, N. Hamedani, and M. R. Ejtehadi, http://xxx.lanl.gov/abs/cond-mat/9905158 (1999).
[56] E. L. Kussell and E. I. Shakhnovich, http://xxx.lanl.gov/abs/cond-mat/9904377 (1999).
[57] H. S. Chan and K. A. Dill, Phys. Today **46**, 24 (1993).
[58] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
[59] H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).
[60] M. A. Roseman, J. Mol. Biol. **200**, 513 (1988).
[61] V. S. Pande, A. Y. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).
[62] S. Borman, Chem. Eng. News **74**, 36 (1996).
[63] F. H. Arnold, Acc. Chem. Res. **31**, 125 (1998).