RESEARCH ARTICLE | NOVEMBER 15 2024

Designing proteins: Mimicking natural protein sequence heterogeneity

Special Collection: Monte Carlo methods, 70 years after Metropolis et al. (1953)

Marcos Lequerica-Mateos ; Jonathan Martin ; José N. Onuchic ; Faruck Morcos ; Ivan Coluzza

J. Chem. Phys. 161, 194102 (2024)
https://doi.org/10.1063/5.0232831
CHORUS

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Designing proteins: Mimicking natural protein sequence heterogeneity

Cite as: J. Chem. Phys. 161, 194102 (2024); doi: 10.1063/5.0232831
Submitted: 10 August 2024 • Accepted: 18 October 2024 •
Published Online: 15 November 2024

Marcos Lequerica-Mateos,[1] Jonathan Martin,[2] José N. Onuchic,[3,4,a)] Faruck Morcos,[2,5,b)] and Ivan Coluzza[3,c)]

AFFILIATIONS

[1] Fundación BCMaterials, UPV/EHU, Leioa, Spain
[2] Department of Biological Sciences, University of Texas at Dallas, Richardson, Texas 75080, USA
[3] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA
[4] Department of Physics and Astronomy, Department of Chemistry, Department of BioSciences, Rice University, Houston, Texas 77005, USA
[5] Departments of Bioengineering, Physics and Center for Systems Biology, University of Texas at Dallas, Richardson, Texas 75080, USA

Note: This paper is part of the JCP Special Topic on Monte Carlo methods, 70 years after Metropolis *et al.* (1953).
[a)] **Electronic mail:** jonuchic@rice.edu
[b)] **Electronic mail:** faruckm@utdallas.edu
[c)] **Author to whom correspondence should be addressed:** ic33@rice.edu

ABSTRACT

This study presents an enhanced protein design algorithm that aims to emulate natural heterogeneity of protein sequences. Initial analysis revealed that natural proteins exhibit a permutation composition lower than the theoretical maximum, suggesting a selective utilization of the 20-letter amino acid alphabet. By not constraining the amino acid composition of the protein sequence but instead allowing random reshuffling of the composition, the resulting design algorithm generates sequences that maintain lower permutation compositions in equilibrium, aligning closely with natural proteins. Folding free energy computations demonstrated that the designed sequences refold to their native structures with high precision, except for proteins with large disordered regions. In addition, direct coupling analysis showed a strong correlation between predicted and actual protein contacts, with accuracy exceeding 82% for a large number of top pairs (>4L). The algorithm also resolved biases in previous designs, ensuring a more accurate representation of protein interactions. Overall, it not only mimics the natural heterogeneity of proteins but also ensures correct folding, marking a significant advancement in protein design and engineering.

I. INTRODUCTION

Protein design is a rapidly growing field with the potential to revolutionize medicine, biotechnology, and our understanding of the principles of life.[1,2] The ability to create proteins with specific functions and properties has profound implications, including the development of highly specific therapeutics and diagnostics for currently incurable diseases, as well as the creation of bio-materials and enzymes with tailored properties, such as environmental responsiveness and enhanced sustainability for the chemical industry, supporting the green revolution.

The field of protein design has advanced significantly in recent years, driven by improvements in computational power, algorithm development, and our understanding of protein structure and function. Current state-of-the-art approaches include both rational design and directed evolution techniques. Rational design requires knowledge of protein structures to engineer proteins with desired structures or functions. This can include designing proteins from scratch with little resemblance to the natural ones or modifying existing proteins to enhance their stability, activity, or specificity. Directed evolution mimics the natural evolutionary process by generating large libraries of protein variants and selecting those with

the desired traits.[3] In addition, machine learning and artificial intelligence are increasingly integrated into protein design workflows, enabling the prediction of protein structures and functions from amino acid sequences with unprecedented accuracy. Recent breakthroughs include the use of deep learning models, such as AlphaFold, which have achieved remarkable success in predicting protein structures with atomic-level precision, thereby providing invaluable insights into the design of proteins.[4]

Despite these advancements, current design approaches have limitations. Rational design often requires extensive knowledge of the target protein structure, which may not always be available. Moreover, it can be challenging to accurately predict the functional consequences of specific modifications. Directed evolution, while powerful, is inherently random and requires high-throughput screening methods that can be time-consuming and resource-intensive. In addition, both methods can struggle with the vast combinatorial space of possible protein sequences, making it difficult to explore all potential designs comprehensively. These limitations highlight the need for more efficient and unbiased methodologies that can explore a wider range of sequence space and yield functional proteins with high precision.

A potential solution to these limitations is the use of coarse-grained protein models for design. These models enable rapid exploration of the sequence space through simple point mutations and residue pair-swapping by simplifying the representation of proteins, thereby significantly accelerating computational processes. However, the use of coarse-grained models comes with its own set of challenges. One significant issue is that these models can easily produce trivial solutions resembling homopolymers, where sequences consist of repetitions of the same few amino acids. Such sequences contain very little information and cannot fold back to a target structure. For instance, various design methods, such as Rosetta, incorporate solutions to avoid homopolymer sequences by using the average composition of amino acids in nature and other constraints to ensure sequence diversity.[5] Previously, it was shown that highly heterogeneous sequences fold systematically better.[6,7] However, when heterogeneity is enforced either through a bias or by fixing the protein composition, the resulting sequences do not exhibit the same level of heterogeneity as natural sequences (see Fig. 1). This discrepancy often leads to designs that cannot be compared to the alignments of natural sequences from different organisms, which are the results of evolution to solve the inverse folding problem.

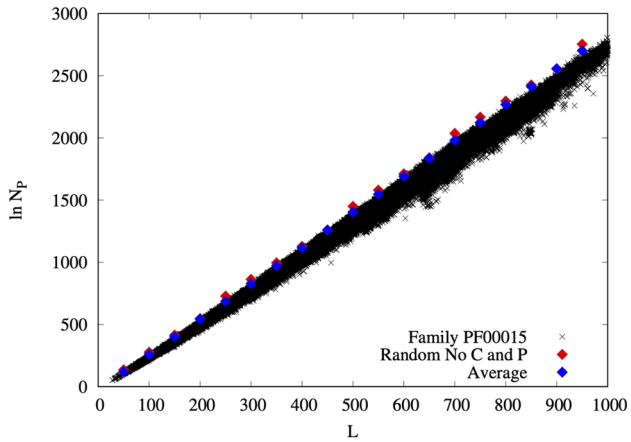

FIG. 1. Logarithm of the number of permutations for a given protein composition vs the protein length. The plot shows three types of proteins: natural proteins from the PF00015 family (black x's), random sequences generated from the average protein composition (blue diamonds), and random sequences excluding cysteine and proline amino acids (red diamonds). Each point represents a protein sequence. The data reveal that natural proteins exhibit a permutation composition lower than the maximum possible with an 18-letter alphabet, and similar to the average composition.

This work presents an approach to protein design that addresses existing limitations by combining point mutation and pair swap moves in a Metropolis Monte Carlo algorithm. This method allows the design of sequences without imposing biases on protein composition, thereby recovering the natural permutations that are lost when average amino acid compositions are used. Such designed sequences are analyzed using direct coupling analysis (DCA), and for specific proteins selected from the pool of design candidates, we test their refolding capabilities. DCA has emerged as a powerful tool for inferring direct interactions between residues within protein sequences based purely on evolutionary data.[8] Our results demonstrate the effectiveness of this approach in designing sequences that fold to a target structure and have compositions compatible with natural proteins.

II. METHODS

A. The Caterpillar protein model

This sequence design procedure is based on the Caterpillar model.[7] This model represents the protein backbone as a fully atomistic five-bead system, with simplified (coarse-grained) side chains depicted by spherically symmetric potentials centered on the Cα atoms. It can simulate both protein folding and the design of foldable proteins, facilitated by the Monte Carlo method used for the simulation.

The Caterpillar model[7] is a coarse-grained model of protein folding that represents the protein backbone using a five-atom representation per residue (see Fig. 1). In this model, the degrees of freedom are limited to the torsional angles ϕ_1 and ϕ_2, while all other structural parameters are kept fixed at values obtained from the literature.[9] This simplified representation allows the model to capture the essential features of protein folding while remaining computationally efficient. The total energy of a protein, E, in the Caterpillar model is given by the sum of the backbone hydrogen bond energy, the side-chain interaction energy, and the residue-solvent interaction energy, scaled appropriately,

$$\begin{aligned} E &= E_H + \alpha E(\varepsilon, E_{\mathrm{HOH}}, \Omega), \\ E(\varepsilon, E_{\mathrm{HOH}}, \Omega) &= \sum_{kl}^{N} \varepsilon_{kl}\Gamma(r_{kl}) + E_{\mathrm{HOH}} \sum_{k}^{N} E_{\mathrm{Sol}}(\Omega - \Omega_k), \end{aligned} \tag{1}$$

where

- E_H is the total energy of the backbone hydrogen bonds as defined in Eq. (2),
- α is a scaling factor that balances the contributions of hydrogen bonding and other interactions, and

- E_{HOH} rescales the Dolittle hydrophobicity index to match the experimental conditions.

Large values of α may lead to the breaking of the maximum valence principle, favoring structures dominated by hydrogen bonds, while small values may overly favor side-chain and solvation interactions, resulting in overly compact structures.

1. Backbone hydrogen bond interactions

Backbone hydrogen bonds are a key factor in stabilizing secondary structures, such as alpha-helices and beta-sheets. In the Caterpillar model, these interactions are modeled using a 10-12 Lennard-Jones type potential,[10] which is defined by the following expression:

$$E_H = -\varepsilon_H(\cos\theta_1\ \cos\theta_2)^{\nu}\left[5\left(\frac{\sigma}{r_{\mathrm{OH}}}\right)^{12} - 6\left(\frac{\sigma}{r_{\mathrm{OH}}}\right)^{10}\right], \tag{2}$$

where

- r_{OH} is the distance between the hydrogen atom of the amide group (NH) and the oxygen atom of the carboxyl group (CO) of the main chain,
- $\sigma = 2.0$ Å is the characteristic distance for hydrogen bonding,
- $\varepsilon_H = -3.1 k_B T_{\mathrm{Ref}}$ is the hydrogen bond strength,
- θ_1 and θ_2 are the angles that determine the orientation of the hydrogen bond, and
- $\nu = 2$ is an exponent that modulates the angular dependence.

The 10-12 Lennard-Jones potential is combined with an angular term $(\cos\theta_1\ \cos\theta_2)^{\nu}$ to ensure proper alignment of the hydrogen bond donor and acceptor groups, stabilizing secondary structures.

2. Side chain interactions

The side chain interactions are modeled using effective sphere–sphere potentials centered on the C_α atoms of the amino acids. These interactions are represented by a sigmoidal function that smoothly transitions from attractive to repulsive regions. The side chain interaction energy between residues i and j is given by

$$E_{ij}(r_{ij}) = S^{ij}_{\mathrm{CAT}}\Gamma(r_{ij}) = S^{ij}_{\mathrm{CAT}}\frac{1}{1+\mathrm{e}^{-(r_{\max}-r_{ij})/W}}, \tag{3}$$

where

- r_{ij} is the distance between the C_α atoms at the centers of spheres i and j;
- $r_{\max} = 12$ Å is the distance at which the potential equals half of its maximum value, $\varepsilon_{ij}/2$;
- $W = 0.4$ Å controls the sharpness of the sigmoidal transition; and
- S^{ij}_{CAT} are elements of a 20 × 20 matrix, each defining the interaction strength between different types of amino acids (see Table S1).

The selection of terms for the S_{CAT} matrix is arbitrary and has varied across different versions of the model. In the initial implementation of the Caterpillar model,[7] the classical Miyazawa–Jernigan potentials[11] and other approaches were tested, all of which successfully led to sequences that folded into the target structures during the design process. In a subsequent study,[6] a method was introduced to adjust the S_{CAT} matrix to align the energies of natural sequences with those generated by the Caterpillar model. This method involves optimizing the interaction parameters of the coarse-grained model by iteratively minimizing the difference between the computed and observed properties of natural protein sequences, using the maximum entropy principle (MEP). This approach ensures that the model parameters produce an interaction matrix closely matching the energy profiles and structural features of natural sequences, leading to a more accurate representation of protein folding and stability. The results were promising, and this version of the S_{CAT} matrix is employed in the current work. Although the current S_{CAT} model captures essential aspects of protein folding, it still lacks the ability to fully reflect the intricate and diverse interactions between amino acids that occur in natural proteins. In natural proteins, amino acid interactions are influenced by various factors, such as side-chain flexibility, chemical environments, and long-range interactions that are highly context-dependent. The current implementation of S_{CAT} simplifies or averages these complexities, meaning it may miss some of the fine details that are critical for accurately representing the full range of amino acid behavior in natural biological systems. It is important to note that, although the amino acids in this model do not directly correspond to biological ones, this virtual system retains the complexity of the design problem and provides valuable insights into the underlying processes. Ongoing work aims to further refine the S_{CAT} matrix to bridge the gap between synthetic Caterpillar sequences and natural ones. This model is shown schematically in Fig. 2(a).

3. Residue–solvent interactions

The residue–solvent interactions are modeled to penalize the exposure of hydrophobic residues and the burial of hydrophilic ones, simulating the effect of solvation on protein stability. This is implemented as a simple energy penalty based on the degree of surface exposure of each residue,

$$E_{\mathrm{Sol}}(\Omega-\Omega_i) = \begin{cases} \varepsilon^i_{\mathrm{Sol}}[\Omega-\Omega_i], & \text{if } \Omega_i \lesseqgtr \Omega,\ \varepsilon^i_{\mathrm{Sol}} \gtrless 0, \\ 0, & \text{if } \Omega_i \gtrless \Omega,\ \varepsilon^i_{\mathrm{Sol}} \gtrless 0, \end{cases} \tag{4}$$

where

- $\Omega_i = \sum_j \Gamma(r_{ij})$ represents the number of contacts that a residue i makes with other residues;
- Ω is the threshold for the number of contacts in the native structure, above which the amino acid is considered fully buried; and
- $\varepsilon^i_{\mathrm{Sol}}$ are the residue-specific solvation energies, taken from the Dolittle hydrophobicity index.[12] They are positive for hydrophobic residues and negative for hydrophilic ones.

This term penalizes the exposure of hydrophobic residues and the burial of hydrophilic ones, thereby promoting a protein structure with a hydrophobic core and a hydrophilic exterior.

B. Virtual move parallel tempering for sequence space sampling

The complex energy landscape associated with protein folding and design requires advanced sampling schemes. To sample

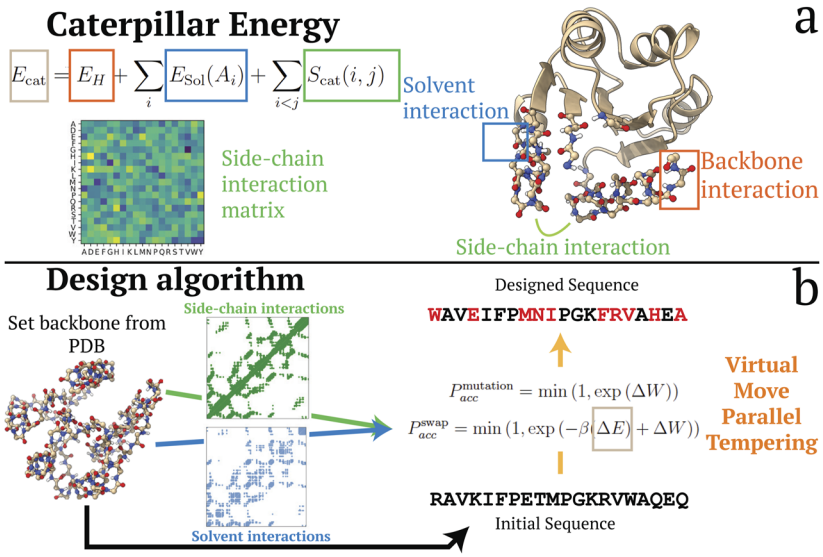

FIG. 2. Schematics of Caterpillar energy and design algorithm used in this work. Caterpillar energy terms are highlighted in a PDB structure in panel (a), with the S_{cat} shown visually below the E_{cat} equation. In panel (b), an overview of the design algorithm is shown.

the sequence space extensively, the virtual move parallel tempering (VMPT) algorithm was employed.[13] In VMPT, the sampling process is distinct from the acceptance of moves. The free energies associated with the collective variables sampled during the simulations are computed from histograms that accumulate the acceptance probabilities of the new and old configurations, respectively. These are the acceptance probability $T_{o\to n}$ for moving from the old to the new state and $T_{n\to o}$ for rejection. The probability $T_{o\to n}$ of accepting a move is given by

$$T_{o\to n} = \frac{\exp(-\beta(\Delta E))}{1+\exp(-\beta(\Delta E))}, \tag{5}$$

while the rejection probability for the old state is

$$T_{n\to o} = \frac{1}{1+\exp(-\beta(\Delta E))}. \tag{6}$$

This approach also accumulates a bias potential W, which adaptively pushes the simulation away from oversampled regions. Over time, W converges to the inverse of the free energy projected over the same collective variable for which the bias potential W is defined.

Multiple replicas at different temperatures are simulated in parallel, with periodic exchanges of configurations governed by the Metropolis criterion, ensuring detailed balance. The actual Markov chain of the Monte Carlo simulation follows the standard Metropolis algorithm, where a trial move is accepted or rejected with probability $P_{acc} = \min(1, \exp(-\beta(\Delta E)))$. Generally, the sampling and acceptance criteria do not have to be the same as long as both respect detailed balance.[14,15]

This decoupling has several advantages. It ensures unbiased sampling, as all configurations are visited with the correct Boltzmann weight. Multiple replicas at different temperatures are simulated in parallel, with periodic exchanges of configurations governed by the Metropolis criterion, and "virtually" exchanged between all temperatures, ensuring uniform sampling across all temperatures and faster convergence of the adaptive bias potential. This, in turn, helps toward more efficient sampling of challenging regions of the free energy landscape since the bias potential can efficiently steer the simulation away from oversampled regions. We provide a visual schematic for this design process in Fig. 2(b).

C. Sequence design

During the design, the backbone of the protein is maintained rigidly while the sequence of the protein is altered using two genetic mutation moves: point mutation (P^{mutation}) and pair swap (P^{swap}). Point mutation generates random variations in sequence composition, while pair swap follows the Metropolis algorithm to minimize the sequence's energy at low temperatures. These operators strike a balance between exploration and exploitation during the optimization process. P^{mutation} introduces stochastic variations into the sequence composition, generating diverse and potentially promising solutions. This expands the search space and prevents premature convergence. In contrast, P^{swap}, inspired by the Metropolis algorithm, optimizes the sequence by minimizing its energy based on the target structure. At low temperatures, it meticulously evaluates the energetic consequences of each swap operation, selecting swaps that reduce the overall energy of the sequence.

The acceptance rules in the original algorithm are

$$P_{\text{acc}}^{\text{mutation}} = \min(1, \exp(-\beta(\Delta E - E_P \ln(\Delta N_P)) + \Delta W)), \tag{7}$$

$$P_{\text{acc}}^{\text{swap}} = \min(1, \exp(-\beta(\Delta E) + \Delta W)), \tag{8}$$

where β is the inverse temperature, ΔE is the energy difference between new and old sequences, and ΔW is the adaptive VMPT bias. The acceptance rule for point mutation includes a bias term $E_P \ln(\Delta N_P)$, where $N_P = N!/(n_a! n_b! \ldots n_z!)$. Here, N is the protein length and $n_a, n_b, \ldots, n_z$ are the numbers of each type of amino acid used. This term acts as a bias forcing the sampling away from homopolymers and toward highly heterogeneous sequences. The

weight $E_P = 10$ sets the relative importance of this bias compared to the Caterpillar energies.

If instead the algorithm is designed to no longer rely on the bias over heterogeneity, the acceptance rules are

$$P_{\text{acc}}^{\text{mutation}} = \min\left(1, \exp\left(\Delta W\right)\right), \tag{9}$$

$$P_{\text{acc}}^{\text{swap}} = \min\left(1, \exp\left(-\beta(\Delta E) + \Delta W\right)\right). \tag{10}$$

In this algorithm, $P_{\text{acc}}^{\text{swap}}$ remains unchanged, while $P_{\text{acc}}^{\text{mutation}}$ no longer depends on the $\Delta E - E_P \ln(\Delta N_P)$ terms, making the Markov chain preferentially sample regions of the sequence space close to random sequences. As we describe in Sec. III, this new mutation probability means that mutations will always be accepted and is a key modification of the algorithm.

However, it is important to note that in VMPT, the sampling process is distinct from the acceptance of moves and is unchanged compared to the previous design algorithm. Thus, the weight for P^{mutation} remains unchanged and will correctly weight the sampled sequences,

$$\text{acc}_{o\to n} = \frac{\exp\left(-\beta(\Delta E - E_P \ln\left(\Delta N_P\right)) + \Delta W\right)}{1 + \exp\left(-\beta(\Delta E - E_P \ln\left(\Delta N_P\right)) + \Delta W\right)} \tag{11}$$

and

$$\text{acc}_{n\to o} = \frac{1}{1 + \exp\left(-\beta(\Delta E - E_P \ln\left(\Delta N_P\right)) + \Delta W\right)}. \tag{12}$$

This decoupling has several advantages. It ensures unbiased sampling, as all configurations are visited with the correct Boltzmann weight. It also allows for more efficient sampling of rare events since move acceptance does not affect configuration probability. In addition, it simplifies implementation, allowing independent sampling and acceptance processes. For the current design approach, VMPT enhances sequence sampling. The decoupling of sampling and acceptance ensures correct sampling when combining random point mutation with standard pair swapping.

D. Direct coupling analysis

The mean field approximation of direct coupling analysis (DCA)[8] can be employed to deduce direct interactions between residues within protein sequences based solely on evolutionary sequence data. This method requires multiple sequence alignment (MSA) of a protein family, which is a collection of homologous sequences that mostly fold into a common structure. Essentially, MSA aligns these sequences to highlight similarities and differences at each position in the protein chain.

DCA works by analyzing the statistical correlations between the positions of residues in the aligned sequences and building a covariance matrix $\mathbf{C} = F_{ij}(A_i, A_j) - f_i(A_i) f_j(A_j)$ of the frequencies F_{ij} and f_i of finding a residue of type A_i at position i when residue A_j is at position j. It distinguishes between direct correlations, which suggest that two residues are likely interacting directly within the protein structure, and indirect correlations, which can arise when two residues both interact with a third residue rather than with each other. The primary challenge in DCA is to correctly identify these direct correlations from the observed data.

The mean field approximation (mfDCA) addresses this challenge by focusing on the most critical interactions and filtering out the noise created by indirect correlations. This simplification not only makes the computational process more efficient but also enhances the accuracy of predicting which residues are directly interacting in the protein structure.

The mathematical foundation of DCA is built on a Boltzmann distribution, where the probability of a protein sequence is defined as being proportional to the Hamiltonian,

$$E_{\text{DCA}}(A) = \sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i). \tag{13}$$

In this equation, $A = (A_1, \ldots, A_L)$ represents the amino acid sequence of the protein, where L is the length of the sequence. The term $e_{ij}(A_i, A_j)$ represents the coupling energies between residues A_i and A_j at positions i and j, indicating the strength and nature of their direct interaction. The term $h_i(A_i)$ represents the local field for residue A_i, reflecting how much a particular residue prefers to be at a specific position in the sequence.

At the core of mfDCA is the relationship between the coupling matrix e_{ij} and the covariance matrix $\mathbf{C}$ derived from the MSA. The covariance matrix captures how residues at different positions co-vary across the sequences in the alignment. By inverting this covariance matrix, mfDCA estimates the coupling energies e_{ij}, thereby identifying direct interactions. This relationship is expressed as $e_{ij}(A_i, A_j) \approx \mathbf{C}_{ij}^{-1}(A_i, A_j)$, linking the statistical analysis directly to the physical interactions in the protein structure.

E. DCA on caterpillar sequences

The inherent properties of the Caterpillar model create an ideal setting for testing the capabilities of DCA. This model generates sequences with minimized structural frustration and eliminates the biological pressures of functional constraints, thus enabling the generation of highly reliable statistical data for DCA. The large number of sequences designed with the Caterpillar model, selected specifically for their ability to fold, provides an ideal framework for generating covariance matrices with superior statistical quality. These matrices are free from the confounding biological functions commonly found in natural proteins. In addition, there is a significant parallel between Eqs. (1) and (13): the S_{cat} matrix, which stores interaction energies between residue types, closely resembles the e_{ij} terms in the DCA Hamiltonian that capture interaction energies between residues at specific positions. The similarity between the energy functional structures of the DCA and Caterpillar energy functions suggests that DCA should be particularly effective when applied to sequences generated by using the Caterpillar design model. Since both models share similar underlying principles in capturing amino acid interactions, we expect DCA to provide highly accurate predictions when analyzing sequences designed by the Caterpillar model. Consequently, the relationship between these terms can be expressed mathematically by

$$S_{\text{cat}}(i,j) \propto \frac{e_{ij}}{\Gamma(r_{ij})} + const, \tag{14}$$

where $\Gamma(r_{ij})$ is the caterpillar radial function for the distance between those positions in the native structure [see Eq. (3)] and *const* is an arbitrary constant.

By averaging e_{ij} for each pair of positions and normalizing by the Caterpillar radial function for the distance between those positions in the native structure, S_{dca} can be inferred, which should be correlated with the S_{cat} used for the design.

F. Analytical solution for the matrix extraction procedure

To analytically identify the relationship between the residue–residue interaction matrix **S** and the DCA covariance matrix **C**, Mathematica 14 was used to invert the covariance matrix **C** and compute the couplings $e_{ij}(A_i, A_j) = -\mathbf{C}^{-1}$ between residues of type A and B at positions i and j, respectively. Since the numerical inversion of **C** is only feasible when sufficient statistical data are available, a numerical approximation is often employed, where a baseline of pseudo-weights pw is uniformly added to **C**.

To this end, it is assumed that all the covariance elements are equal except for the block diagonal terms corresponding to the homo-residue interactions, which are scaled by a factor $b < 1$ to imitate a lower probability of encountering homo-pairs.

The protein length is N while the alphabet size including gaps is $q = 21$ and includes the gaps. Corrections to the covariance matrix $C = FPW_{ij}^{\alpha\beta} - f_{pw}(i\alpha) f_{pw}(j\beta)$, where p_w is the pseudo-count weight and where

$$FPW_{ii}^{\alpha\alpha} = b(1-p_w)F_{ii}^{\alpha\alpha} + \frac{p_w}{q}, \quad FPW_{ii}^{\alpha\beta} = (1-p_w)F_{ii}^{\alpha\beta} + \frac{p_w}{q}, \tag{15}$$

$$FPW_{ij}^{\alpha\alpha} = b(1-p_w)F_{ij}^{\alpha\alpha} + \frac{p_w}{q^2}; j \neq i, FPW_{ij}^{\alpha\beta} = (1-p_w)F_{ij}^{\alpha\beta} + \frac{p_w}{q^2}; \quad \alpha \neq \beta; j \neq i, \tag{16}$$

$$f_{pw}^{i\alpha} = (1-p_w) f_{i\alpha} + \frac{p_w}{q}. \tag{17}$$

Here, we indicated with α and β the amino acids types from the 20 letter alphabet, while i and j are the residue indices along the sequence. Assuming that $F_{ij}^{\alpha\beta} = F$ and $f_{i\alpha} = f$, the pseudo-weighted covariance matrix C is then

$$\mathbf{C} = \begin{pmatrix} \begin{bmatrix} b(1-p_w)F + \frac{p_w}{q} & \cdots & (1-p_w)F \\ \vdots & \ddots & \vdots \\ (1-p_w)F & \cdots & b(1-p_w)F + \frac{p_w}{q} \end{bmatrix} & \cdots & \begin{bmatrix} b(1-p_w)F + \frac{p_w}{q^2} & \cdots & (1-p_w)F + \frac{p_w}{q^2} \\ \vdots & \ddots & \vdots \\ (1-p_w)F + \frac{p_w}{q^2} & \cdots & b(1-p_w)F + \frac{p_w}{q^2} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} b(1-p_w)F + \frac{p_w}{q^2} & \cdots & (1-p_w)F + \frac{p_w}{q^2} \\ \vdots & \ddots & \vdots \\ (1-p_w)F + \frac{p_w}{q^2} & \cdots & b(1-p_w)F + \frac{p_w}{q^2} \end{bmatrix} & \cdots & \begin{bmatrix} b(1-p_w)F + \frac{p_w}{q} & \cdots & (1-p_w)F \\ \vdots & \ddots & \vdots \\ (1-p_w)F & \cdots & b(1-p_w)F + \frac{p_w}{q} \end{bmatrix} \end{pmatrix} - \begin{pmatrix} \left((1-p_w)f + \frac{p_w}{q}\right)^2 & \cdots & \left((1-p_w)f + \frac{p_w}{q}\right)^2 \\ \vdots & \ddots & \vdots \\ \left((1-p_w)f + \frac{p_w}{q}\right)^2 & \cdots & \left((1-p_w)f + \frac{p_w}{q}\right)^2 \end{pmatrix}, \tag{18}$$

where there are N blocks and each block is of size $q - 1$.

The interaction matrix **S** is then obtained by accumulating elements from $\mathbf{C}^{-1}$ based on certain conditions,

$$\mathbf{S}[k,l] = \sum_{\substack{i,j \\ \mathrm{abs}(i-j)>4}} \mathbf{C}^{-1}[i\cdot(q-1)+k, j\cdot(q-1)+l]. \tag{19}$$

By solving Eq. (19) numerically, it is found that the **S** matrix has one value on the diagonal $\mathbf{S}[k,k] = \mathbf{S}_{\mathrm{diag}}$ and another for all the elements off the diagonal $\mathbf{S}[k,l] = \mathbf{S}_{\mathrm{off\,diag}}$; $k \neq l$. The difference between them, $G = \mathbf{S}_{\mathrm{diag}} - \mathbf{S}_{\mathrm{off\,diag}}$, results in

$$G \propto \frac{(b-1)}{b+1}, \tag{20}$$

which for $1 > b > 0$ results in a flat **S** matrix with a positive diagonal shift of G; a direct result of homo-residue pairs have a systematic lower probability of appearing controlled by the factor b.

G. Folding simulations

To characterize the equilibrium configuration of each design, we selected on sequence per proteins (for a total of four sequences) to test their refolding properties. During the folding simulation, the folding free energy F is computed as a function of several order

parameters. Each case will be described separately later in this article, but all will be similar to the following example.

First, a convenient order parameter that tracks the progress of the folding process is defined. The chosen parameter is the distance root mean square deviation (DRMSD). The DRMSD is a metric used to quantify the difference between two sets of atomic coordinates, typically in the context of comparing the structural similarity of proteins or other macromolecules. DRMSD is calculated by first determining the pairwise distances between equivalent atoms in the two structures being compared. These distances are then squared, summed, and averaged over all atom pairs. Finally, the square root of this average is taken to obtain the DRMSD value. Mathematically, DRMSD is defined as

$$\mathrm{DRMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - d_i')^2},$$

where N is the total number of atom pairs, d_i is the distance between the ith pair of atoms in the first structure, and d_i' is the corresponding distance in the second structure. DRMSD provides a comprehensive measure of structural deviation, making it a useful tool for assessing the degree of conformational changes or structural similarity between macromolecular models.

To compute the free energy $F(\mathrm{DRMSD})$, the following relation is used:

$$F(DRMSD) = -kT \ln[P(DRMSD)], \tag{21}$$

where $F(\mathrm{DRMSD})$ is the free energy of the state with the order parameter DRMSD and $P(\mathrm{DRMSD})$ denotes a normalized histogram of the number of sampled conformations with the order parameter DRMSD. In practice, a direct (brute force) calculation of this histogram is inefficient, as the system tends to become trapped in local minima, especially at low temperatures. To address this issue, the virtual move parallel tempering (VMPT) algorithm is incorporated.

III. RESULTS AND DISCUSSION

Initially, we analyzed the permutations from the PF00015 family, one of the largest protein families, as a representative of natural proteins and studied them as a function of their length (see Fig. 1). The results demonstrate that natural proteins have a permutation composition lower than the maximum possible with an 18-letter alphabet. This suggests that natural proteins, on average, do not utilize the full extent of the 20-letter alphabet. Even with a reduced alphabet, it would be atypical to bias toward the highest permutation values. Historically, average compositions from natural proteins

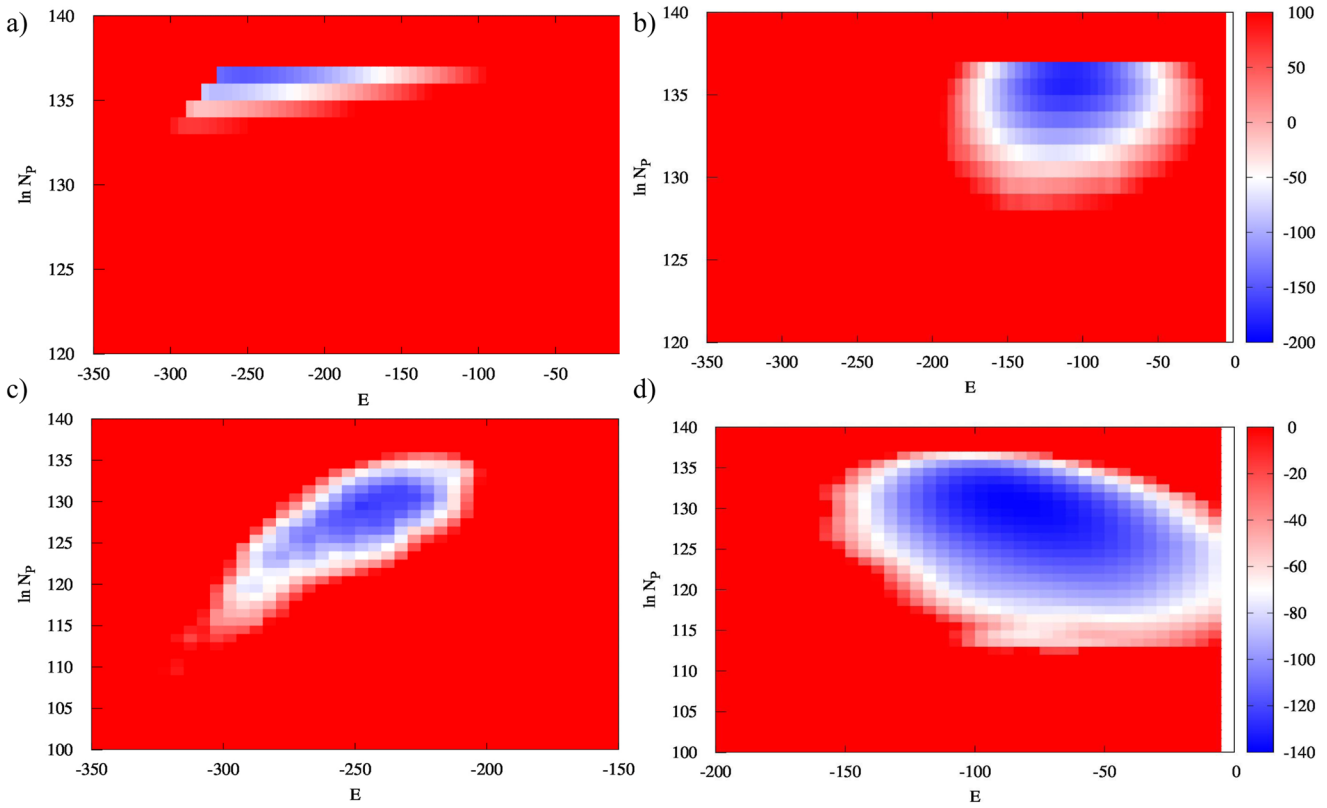

FIG. 3. Plot of ln N_P vs energy E for the (a) low temperature T = 0.5 (reduced units), (b) high temperature T = 10 (reduced units) for the constrained algorithm, (c) low temperature T = 0.1 (reduced units), and (d) high temperature T = 10 (reduced units) for the unconstrained algorithm. On comparing panels (c) and (d), it is visible that the unconstrained design algorithm produces sequences with the same heterogeneity ln N_P at low and high temperature.

have been used as constraints for design. Figure 1 shows no significant difference between the two approaches, but both yield sequences with permutations higher than natural ones.

Given that both traditional methods and the Caterpillar design exhibit heterogeneity systematically larger than that of natural sequences, an alternative method is proposed. In addition, in a previous study,[6,16] it was observed that sequences generated with the design algorithm had a systematically lower probability of homo-pairs. This effect resulted from the heterogeneity $\ln N_P$, which slightly lowers the probability of homo-residue pairs (e.g., AA or LL), leading, incorrectly, to more repulsive interactions along the matrix diagonal. To further support this hypothesis, the DCA inversion procedure for a covariance matrix with all equal values except for terms corresponding to homo-residue pairs was analytically solved, applying a small shift to lower the probability (see Sec. II). If an effective interaction is extracted from such a covariance matrix, a flat interaction matrix with a positive diagonal is obtained. If instead $b = 0$, the diagonal term in the extracted interaction matrix disappears.

This observation led us to substitute the bias over permutations with a point mutation move that is always accepted, thereby allowing the algorithm to run without bias at high temperature (see Sec. II). This modification aimed to produce sequences more closely mirroring the natural heterogeneity observed in proteins. In the following sections, the previous design approach will be referred to as *constrained*, in contrast to the current approach, which will be referred to as *unconstrained*.

Designed sequences were generated using the unconstrained algorithm. The sequences in equilibrium exhibit lower composition permutations. This change in equilibrium composition is apparent when projecting the sequences sampled during the design simulation over the free energy landscape $F(E, \ln N_P)$, which is a function of two collective variables: the energy E and the heterogeneity $\ln N_P$. By comparing $F(E, \ln N_P)$ calculated with the constrained design approach and the unconstrained one (Fig. 3), it is observed that the design methods at a low temperature ($T = 0.1$ in reduced units) achieve similar equilibrium energies but very different heterogeneity $\ln N_P$.

The unconstrained approach yields compositions compatible with natural proteins regarding the number of permutations (see Fig. 4). On the other hand, the similarity between the equilibrium energies indicates a similar level of optimization.

A. Refolding tests of sequences generated with the unconstrained design algorithm

Having demonstrated that the unconstrained design algorithm leads to sequences with more natural heterogeneity, tests were conducted to determine if the sequences still fold back to the target structures. To achieve this, a sequence for each of the target proteins corresponding to the minimum of the free energy $F(E, \ln N_P)$ at the lowest temperature $T = 0.1$ was isolated. For each of the isolated sequences, the folding free energy $F(\mathrm{DRMSD})$ was computed (see *Methods* for details), during which the protein conformational landscape is projected over the DRMSD from the native state. In the past, this combination has proven effective in estimating the folding ability of artificial sequences.[17] The computed landscapes (see Fig. 5) showed consistent refolding within 2 and 4 Å from the native structure. An exception is the FKBP12 protein (PDBID: 2PPN), which has a large disordered region. Ignoring that region in the RMSD calculations, the folded core of the protein exhibits a refolded structure with similar precision as the other structures (see 3D comparison in Fig. S1).

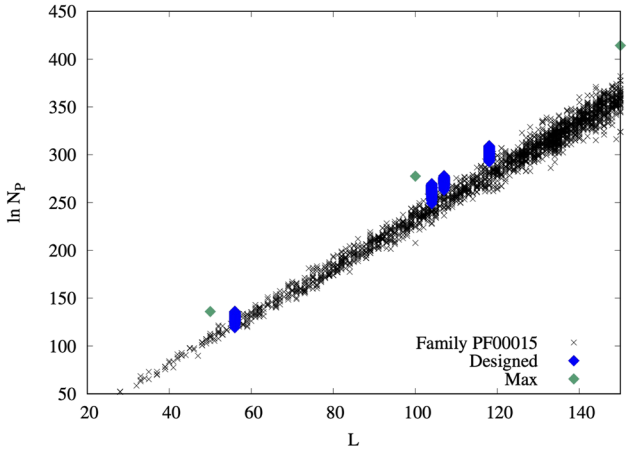

FIG. 4. Logarithm of the number of permutations for a given protein composition vs protein length. The plot shows three types of proteins: natural proteins from the methyl-accepting chemotaxis protein (MCP) signaling family (PF00015, black x's), artificially designed proteins (blue diamonds), and random sequences excluding cysteine and proline amino acids (green diamonds). Each point represents a protein sequence. The plot demonstrates that the unconstrained design approach produces compositions compatible with natural proteins regarding the number of permutations.

To assess the quality of all designed sequences, rather than just a selected few, DCA was performed on all the sequences sampled in the low-temperature free energy minima. First, the direct

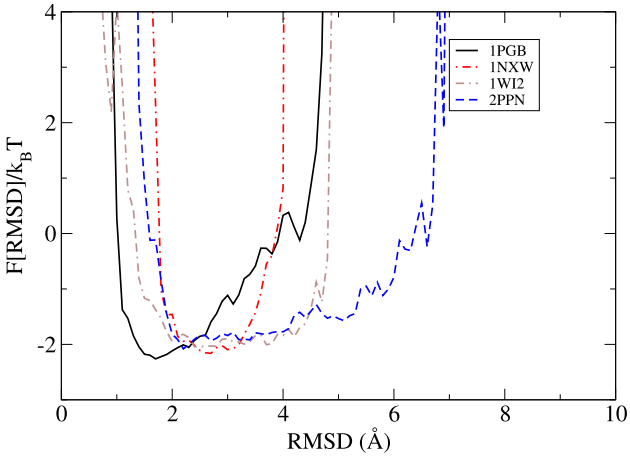

FIG. 5. Refolding tests of sequences generated with the unconstrained design algorithm. Refolding of the sequences with the corresponding free energy calculation of the conformational landscape projected over the DRMSD from the native state showed consistent refolding within 2–4 Å from the native structure. The exception is the FKBP12 protein (PDBID: 2PPN), which has a large disordered region. If that region is ignored in the RMSD calculations, the folded core of the protein shows a refolded structure with similar precision as the other structures.

information (DI) pairs were computed, for which the top ones should, theoretically, have a high correlation with the actual contacts between residues. In fact, DI pairs are used as an effective contact prediction approach for structure prediction.[18] When applied to sequences designed with the constrained algorithm to fold to a target structure, the agreement with DI and actual protein contacts is above 95% for the top 500 pairs. Comparable accuracy is achieved with the current design algorithm for the proteins 2PPN, 1NXW, 1PGB, and 1WI2 (see Fig. 6).

The final test involves extracting the interaction matrices from the direct coupling interactions by averaging the DI over all the protein contacts (see Sec. II), as described in a recent

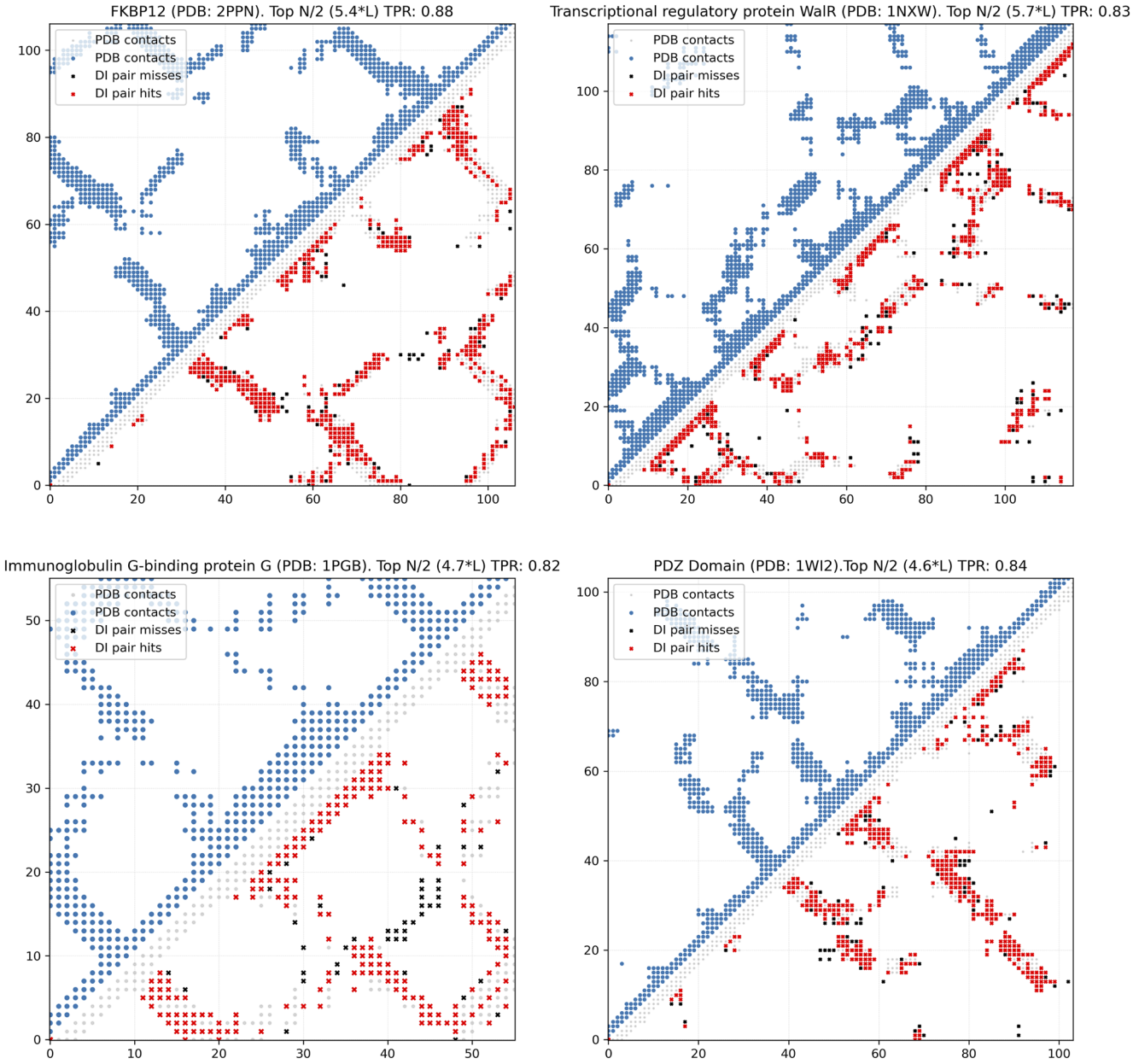

FIG. 6. Comparison of predicted and crystallographic contacts for four proteins. True positive contacts predicted from DCA of artificial sequences generated using the unconstrained design algorithm are represented in red, and false positive contacts are represented by the black crosses. Actual contacts determined from crystallography are represented in blue and gray (contacts defined as α-carbon atoms within 12 Å of each other). The figure demonstrates an excellent agreement between the predicted and crystallographic contacts.

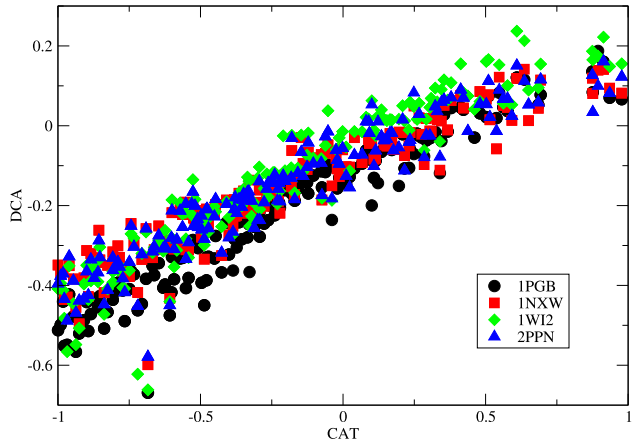

FIG. 7. Scatterplot comparing residue–residue interaction matrices computed using two different methods, Caterpillar and DCA, for four proteins of varying lengths. Each point represents a residue pair from the proteins 1PGB (56 residues), 1NXW (118 residues), 1WI2 (104 residues), and 2PPN (107 residues). The x axis shows the Caterpillar interaction values, while the y axis shows the DCA interaction values. The correlation between the two methods is evident from the clustering of points along the diagonal line. The white band in the plot at (0.8, 0.1) is due to the exclusion of cysteine and proline terms in the *S* matrix.

publication using the previous design approach.[19] The previous results indicated that the original interaction matrix can be extracted from a family of sequences designed to fold into a target structure. Figure 7 shows that the same result is reproduced with the updated design approach. Notably, for the updated algorithm, the gap between the diagonal and the other elements of the DCA couplings matrix is no longer present (see Fig. 7).

The observed strong correlation between direct information (DI) predictions and the protein contact map is noteworthy. This accurate prediction of contacts serves as an initial indication that these sequences are viable solutions to the protein folding problem. Predicting a high number of true positive contacts implies that the equilibrium configuration is likely the folded structure. If this principle did not hold, structure-based models would not function as effectively as they do.[20] Furthermore, in cases where there is no well-defined structure to fold into, such as with random proteins or intrinsically disordered proteins (IDPs), DCA produces only a faint signal, reinforcing this concept.[21]

IV. CONCLUSIONS

This study explored the design and analysis of protein sequences using an algorithm that aims to match the heterogeneity of natural sequences. Our initial analysis demonstrated that natural proteins have a permutation composition lower than the theoretical maximum with an 18-letter alphabet, suggesting they do not utilize the full 20-letter alphabet to its highest permutation values. This finding was consistent with historical data where average compositions from natural proteins were used as design constraints.

An alternative algorithm was designed to generate sequences that, when compared to natural proteins, shows a lower permutation composition. This equilibrium shift was further illustrated through the free energy landscape $F(E, \ln N_P)$, which revealed that while the equilibrium energies were similar between the constrained and unconstrained design methods, the heterogeneity was significantly different. The unconstrained design approach produced sequences with permutation compositions compatible with natural proteins, indicating a refined level of optimization.

To validate the effectiveness of this algorithm, the refolding capabilities of the designed sequences were tested. The folding free energy $F(\mathrm{DRMSD})$ computations confirmed that these sequences could refold to their native structures within 2–4 Å, except for the 2PPN protein, which contains a large disordered region. Ignoring this region, the refolded core structure displayed similar precision to other proteins, underscoring the robustness of the unconstrained design approach.

Furthermore, DCA performed on all designed sequences revealed an excellent correlation between the direct information (DI) pairs and the actual contacts within the target protein structures, with an accuracy above 82% for a large number of top pairs (>4L) for proteins, such as 1NXW, 1PGB, 1WI2, and 2PPN. This high level of agreement suggests that the unconstrained algorithm reliably produces important sequence changes that favor folding and correct contact formation of the target structure, comparable with natural proteins.

Finally, it was demonstrated that interaction matrices derived from the unconstrained design approach showed no artificial gap between the diagonal and other elements, a problem observed in the constrained algorithm due to biases in heterogeneity. This confirms that the unconstrained design approach provides a more accurate representation of protein interactions.

In summary, this algorithm not only produces sequences that mirror the natural heterogeneity of proteins but also generates sequences that can fold correctly, making it a significant advancement in the field of protein design. The results of this study pave the way for more refined and efficient protein engineering methodologies that better mimic the complexity of natural proteins. Furthermore, the algorithm can be extended to design novel folds. In fact, we can leverage the same strategy used in previous studies,[16,22] where we successfully identified novel folds for generalized heteropolymers, including specific knotted topologies. The approach involves an initial prescreening of the folded landscape using an algorithm that we refer to as SEEK, which identifies the best candidates for further design refinement. By applying this methodology, we could explore a wider range of protein folds, advancing the ability to design novel structures with unprecedented complexity and specificity.

SUPPLEMENTARY MATERIAL

The supplementary material includes the following: Figure S1: comparing x-ray and refolded structures using the Caterpillar model to demonstrate foldability prediction accuracy and Fig. S2: showing strong agreement between DCA-predicted and Caterpillar refolded contacts for the studied proteins. These materials provide additional validation for the findings in the main publication.

ACKNOWLEDGMENTS

The work at the Center for Theoretical Biological Physics was sponsored by the National Science Foundation (Grant Nos.

PHY-2019745 and PHY-2210291) and by the Welch Foundation (Grant No. C-1792) J.N.O. is a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas. J.M. and F.M. acknowledge the support from the National Institutes of Health (NIH No. R35GM133631). F.M. acknowledges the support from the NSF CAREER award (No. MCB-1943442).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Marcos Lequerica-Mateos: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Jonathan Martin**: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **José N. Onuchic**: Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Faruck Morcos**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ivan Coluzza**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] P.-S. Huang, S. E. Boyken, and D. Baker, Nature **537**, 320 (2016).
[2] E. Marcos and D. Silva, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**, e1374 (2018).
[3] L. Sellés Vidal, M. Isalan, J. T. Heap, and R. Ledesma-Amaro, RSC Chem. Biol. **4**, 271 (2023).
[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Nature **596**, 583 (2021).
[5] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. J. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray, J. Chem. Theory Comput. **13**, 3031 (2017).
[6] I. Coluzza, PLoS One **9**, e112852 (2014).
[7] I. Coluzza, PLoS One **6**, e20853 (2011).
[8] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U. S. A. **108**, E1293 (2011).
[9] R. Freedman, Trends Biochem. Sci. **10**, 82 (1985).
[10] A. Irbäck, F. Sjunnesson, and S. Wallin, Proc. Natl. Acad. Sci. U. S. A. **97**, 13614 (2000); arXiv:cond-mat/0011079.
[11] S. Miyazawa, L. R. Jernigan, R. L. Jernigan, M. Biology, N. Institutes, L. R. Jernigan, R. L. Jernigan, M. Biology, and N. Institutes, Macromolecules **18**, 534 (1985).
[12] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[13] I. Coluzza and D. Frenkel, ChemPhysChem **6**, 1779 (2005); arXiv:cond-mat/0503245.
[14] G. C. Boulougouris and D. Frenkel, J. Chem. Phys. **122**, 244106 (2005).
[15] G. C. Boulougouris and D. Frenkel, J. Chem. Theory Comput. **1**, 389 (2005).
[16] C. Cardelli, F. Nerattini, L. Tubiana, V. Bianco, C. Dellago, F. Sciortino, and I. Coluzza, Adv. Theory Simul. **2**, 1900031 (2019).
[17] I. Coluzza, Mapping Ignorance (2019).
[18] J. I. Sulkowska, J. K. Noel, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **109**, 17783 (2012).
[19] J. Martin, M. Lequerica Mateos, J. N. Onuchic, I. Coluzza, and F. Morcos, Proc. Natl. Acad. Sci. U. S. A. **121**, e2311807121 (2024).
[20] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[21] J. A. Iserte, T. Lazar, S. C. E. Tosatto, P. Tompa, and C. Marino-Buslje, Sci. Rep. **10**, 17962 (2020).
[22] I. Coluzza, P. D. J. Van Oostrum, B. Capone, E. Reimhult, and C. Dellago, Phys. Rev. Lett. **110**, 075501 (2013).