Protein Design by Integrating Machine Learning and Quantum-Encoded Optimization

Veronica Panizza,[1,2] Philipp Hauke,[1,2,*] Cristian Micheletti,[3,†] and Pietro Faccioli[4,‡]

[1]*Pitaevskii BEC Center CNR-INO and Physics Department, Trento University, Via Sommarive 14, 38123 Povo (Trento), Italy*

[2]*INFN-TIFPA, Via Sommarive 14, 38123 Povo (Trento), Italy*

[3]*Scuola Internazionale Superiore di Studi Avanzati (SISSA),Via Bonomea 265, I-34136 Trieste, Italy*

[4]*Department of Physics, University of Milano-Bicocca and INFN, Piazza della Scienza 3, I-20126 Milan, Italy*

(Received 6 May 2024; accepted 22 October 2024; published 15 November 2024)

The protein design problem involves finding polypeptide sequences folding into a given three-dimensional structure. Its rigorous algorithmic solution is computationally demanding, involving a nested search in sequence and structure spaces. Structure searches can now be bypassed thanks to recent machine-learning breakthroughs, which have enabled accurate and rapid structure predictions. Similarly, sequence searches might be entirely transformed by the advent of quantum annealing machines and by the required new encodings of the search problem, which could be performative even on classical machines. In this work, we introduce a general protein design scheme where algorithmic and technological advancements in machine learning and quantum-inspired algorithms can be integrated, and an optimal physics-based scoring function is iteratively learned. In this first proof-of-concept application, we apply the iterative method to a lattice protein model amenable to exhaustive benchmarks, finding that it can rapidly learn a physics-based scoring function and achieve promising design performances. Strikingly, our quantum-inspired reformulation outperforms conventional sequence optimization even when adopted on classical machines. The scheme is general and can be extended, e.g., to encompass off-lattice models, and it can integrate progress on various computational platforms, thus representing a new paradigm approach for protein design.

DOI: 10.1103/PRXLife.2.043012

I. INTRODUCTION

In contrast to random polypeptide chains, most naturally occurring proteins fold rapidly and reversibly into a unique conformation that is solely determined by the sequence of amino acids, called the native state [1–3]. This property is consistent with the interpretation that the native state typically corresponds to the free-energy minimum of the peptide chain [4,5] and is kinetically accessible from generic conformers of the unfolded ensemble [2,3,6,7]. The unique thermodynamic properties of proteins and proteinlike systems promoted by natural or artificial selection [2–5,8–18] have long posed two fundamental challenges: (i) predicting protein structures given the chemical sequence, and (ii) finding sequences that can fold onto a given target structure. These are known as protein folding and protein design problems, respectively. Because of their close connection, they are also called the *direct* and the *inverse* protein folding problems.

From a thermodynamic perspective, both challenges can be fully specified by defining the (effective) energy $E(\Gamma, S)$ of a polypeptide chain as a function of its sequence, S, and its conformational state, Γ. By effective energy, we intend that E includes contributions from the thermodynamic integration of the solvent degrees of freedom. Solving the direct protein folding problem for a given polypeptide sequence S involves finding the conformer(s) Γ with the largest occupation probability in canonical equilibrium,

$$P_{\rm opt}(\Gamma|S) = \max_{\Gamma} \frac{e^{-\beta E(\Gamma,S)}}{\sum_{\Gamma'} e^{-\beta E(\Gamma',S)}} \equiv \max_{\Gamma} e^{-\beta(E(\Gamma,S)-F(S))} \geqslant p_{\rm fold}, \tag{1}$$

where β is the inverse thermal energy in physiological conditions, and $F(S) = -\frac{1}{\beta}\ln\sum_{\Gamma'} e^{-\beta E(\Gamma',S)}$ is the free energy of sequence S, which involves the sum over the possible conformational states. Foldable polypeptide chains, such as naturally occurring proteins, are characterized by the thermodynamic stability of the state Γ maximizing Eq. (1), i.e., $P_{\rm opt}(\Gamma|S) > p_{\rm fold}$, where $p_{\rm fold}$ is a suitable threshold, typically larger than 0.5. In this case, Γ is the native state of S.

Conversely, solving the inverse folding problem for a given target state Γ_T amounts to finding a sequence S, if any exists, such that

$$P_{\rm opt}(S|\Gamma_T) = \max_{S} e^{-\beta(E(\Gamma_T,S)-F(S))} \geqslant p_{\rm fold}. \tag{2}$$

Sequences that satisfy inequality (2) are said to design the target state Γ_T [19,20].

*Contact author: philipp.hauke@unitn.it

†Contact author: cristian.micheletti@sissa.it

‡Contact author: pietro.faccioli@unimib.it

Thus, solving the design problem is equivalent to finding sequences that minimize the function

$$G(S) = E(\Gamma_T, S) - F(S), \tag{3}$$

and satisfy the inequality in Eq. (2).

A key point is that the computational demands for solving the above rigorous physics-based formulations of the direct and inverse folding problems differ greatly. Solving the former involves, in principle, computing the energy of the sequence of interest over the entire set of physically viable conformational states. Conversely, to solve the design problem, it is not sufficient to compute the energy of all viable sequences on the given target structure, Γ_T. Indeed, the sequences that minimize the energy when mounted on Γ_T may fold into a different structure, Γ', with even lower effective energy, i.e., $E(\Gamma', S) < E(\Gamma_T, S)$ [19,20]. Hence, solving the design problem involves two nested searches: one over the sequences and one over the structures [19–22]. For this reason, much effort has been spent on finding practical schemes and approximations to curb the computational expenditure entailed by this problem [19–25].

A novel twist in this direction has come from recent advancements in machine learning [26–28]. Relevant examples include the development of Bayesian learning design strategies [29], and the experimental validation of deep-learning models [30], including generative ones [31,32]. On the one hand, these approaches provide elegant and valuable demonstrations of the striking extent to which sequence-structure correlations present in protein databases might be harnessed by empirical scoring functions for protein design. In perspective, such endeavors could emulate the breakthrough in the empirical solution to the direct folding problem, where deep neural networks now yield remarkably reliable predictions [33].

On the other hand, an inherent limitation of all such empirical methods is the lack of interpretability. In contrast, physics-based approaches, based on an explicit definition of the energy function $E(\Gamma, S)$, would enable abstracting principles applicable to more general contexts [34]. For this reason, the quest for computationally amenable physics-based approaches to the protein design and related problems remains an active research avenue as well as a natural testbed for new computing hardware paradigms, including quantum computing [35–39], as we shall discuss later.

In physics-based schemes, the cornerstone notion is that the energy function $E(\Gamma, S)$ is the only theoretical ingredient needed to solve both the direct and the inverse folding problems. However, in detailed atomistic approaches, even a single computation of $E(\Gamma, S)$ would require extensive calculations, e.g., to integrate out the solvent degrees of freedom. Customarily, this prohibitively expensive computation is alleviated by resorting to coarse-grained models and implicit-solvent energy functions. At the same time, coarse-graining also tames the complexity of the sampling problem by smoothing the energy landscape and drastically reducing the number of conformational degrees of freedom [40–45]. Yet, reliably estimating the functional form and the parametrization of the effective energy $E(\Gamma, S)$ remains challenging.

The main goal of the present study is to demonstrate that it is possible to integrate advancements in both machine learning and quantum computing technologies to tackle the design problem without abandoning the physics-based standpoint of Eqs. (2) and (3). In this context, the research in quantum computing may also drive the development of radically new physics-based formulations that are advantageous even when implemented on classical machines [46].

In recent years, several quantum-inspired algorithms have been proposed to unveil protein sequence-structure relationship [35–37,39,47,48], compute protein folding pathways [49,50], and more generally address the equilibrium properties of polymeric systems [46,51]. By contrast, the protein design problem has been tackled by comparatively fewer attempts using algorithms designed for quantum hardware. Such pioneering efforts have relied on lattice protein models [9,10,52,53] because their discrete nature enables a straightforward mapping onto the quantum simulation hardware. In Ref. [38], the authors used a gate-based quantum algorithm to reshuffle a sequence to minimize its energy on a reference structure. In contrast, in Ref. [54] a quantum-annealing platform was used for an analogous objective. Both studies employed a simplified two-letter amino acid alphabet and postulated the effective amino acid interactions.

In our first illustrative application, we also resort to minimalistic lattice protein models. This choice is particularly suited to assessing the accuracy of our scheme, allowing us to better control the sources of errors. Indeed, it eliminates the uncertainties associated with heuristic machine-learning algorithms for protein folding, as they can be replaced by an exhaustive search of the conformational space. Furthermore, it enables us to assess the accuracy through which our iterative learning scheme is able to learn the underlying physics-based energy function.

Even after this major simplification, the combinatorial search over the sequence space can be computationally demanding. Quantum annealing machines are ideally suited to solve this kind of discrete combinatorial optimization after a suitable mathematical reformulation, or encoding, of the original problem. A very relevant question to address is if such reformulation can lead to performance improvement even when adopted on classical machines [46].

We answer this question in the affirmative. Indeed, in our proof-of-concept study, quantum-encoded approaches implemented on both classical and quantum computers outperform a well-established scheme based on simulated annealing. At the same time, our iterative machine-learning scheme enables us to reach solutions to the design problem with a high success rate. Our algorithm's main merit is its ability to simultaneously harness the new possibilities offered by machine-learning applications and promised by the advancements in quantum computing hardware while remaining rooted into the physics-based modeling paradigm. Furthermore, its portability to off-lattice all-atom molecular representations represents an important stepping stone towards perspective realistic applications. Collectively, our results suggest that, if the size and performance of quantum simulators continue to improve over the next several years, the integration of quantum annealers and classical machine learning may represent a transformative new paradigm with prospective implications in several areas of life sciences and pharmacology.

II. MODELS AND METHODS

A. Approximate scoring function $G(S)$

We simplify the complexity of the design problem by introducing two approximations to circumvent the nested sequence-structure search implied by the minimization of the design scoring function in Eq. (3).

First, we resort to a customary linear ansatz for $E(\Gamma, S)$:

$$E(\Gamma, S) \simeq \sum_{i,j=1,\ldots,n}{}' C_{ij}(\Gamma)\, \varepsilon_{s_i,s_j}. \tag{4}$$

Here, s_i is the chemical identity of the ith amino acid of the sequence S, which has length n, ε_{s_i,s_j} are the entries of a suitable $D\times D$ energy matrix—D being the size of the amino acids chemical alphabet—and $C(\Gamma)$ is the contact map of the conformational state Γ, with entries $C(\Gamma)_{ij}$ equal to 1 if amino acids i and j are in contact, and equal to 0 otherwise. The primed summation indicates the restriction to distinct pairs, $j > i$. A weighted contact map, with entries spanning the entire [0:1] interval, could also be used instead of the binary one.

In Eq. (3), the energy of the sequence S mounted over the target structure is computed relative to the sequence free energy $F(S)$. Evaluating the latter implies computing the energy of S mounted over all possible states, a computationally prohibitive task. A key approximation of our approach consists in replacing this reference with the average energy evaluated over a database of known native structures,

$$F(S) \simeq \sum_{i,j=1,\ldots,n}{}' \varepsilon(s_i, s_j)\, \langle C_{ij}\rangle. \tag{5}$$

In the expression above, $\langle C_{ij}\rangle$ is the average contact map of those structures in the database that have the same length as the target one. In more general contexts, this restriction can be relaxed by setting $\langle C_{ij}\rangle$ equal to the contact probability of amino acids at chemical distance $|i-j|$ computed over all database entries that are sufficiently longer than $|i-j|$, to avoid end effects.

The gist of the approximation in Eq. (5) is to yield a free-energy estimate that, while remaining computationally amenable, is still informed by the structural properties of viable states. The average pairwise contact probabilities appear to be the most natural and effective choice in this respect, considering that the approach could be systematically generalized to include three-body and higher-order contact probabilities.

With this proviso, our approximation to the design scoring function $G(S)$ becomes

$$G(S) \simeq \sum_{i,j=1,\ldots,n}{}' \varepsilon(s_i, s_j)\, [C_{ij}(\Gamma_T) - \langle C_{ij}\rangle]. \tag{6}$$

Minimizing this function thus selects sequences whose native energy is as low as possible compared to the average taken over the database. Importantly, prior knowledge of the energy matrix $\varepsilon(s_i, s_j)$ is not needed. Instead, we propose an iterative approach, discussed in the next section, through which an initial guess is refined until consistency between the direct and inverse folding problems is reached.

B. Iterative scheme to tackle the design problem

Key parameters of our design scheme are the entries of the $D\times D$ symmetric energy matrix $\boldsymbol{\varepsilon}$ of Eq. (6). Inspired by earlier work on the extraction of effective potentials for protein folding or design [55–58], we adopted an iterative scheme based on enforcing consistency between the solutions of the direct and inverse folding problems. Our choice is motivated by the increasing availability of reliable and fast algorithms for predicting protein folds even in realistic contexts [33]. In principle, this opens the possibility of harnessing these efficient methods for tuning $\boldsymbol{\varepsilon}$ to design a specific type or family of target structures. Such optimized schemes would also limit adverse effects inherent to structural coarse-graining, which inevitably impacts the transferability of potential energies obtained by thermodynamic integration. For the same reason, it is not apparent *a priori* that database-wide schemes for extracting interaction potentials, including the powerful quasichemical approximation [59–63], are well suited to the design task based on the minimization of Eq. (6).

The key steps of our iterative scheme are sketched in Fig. 1. Given the target structure to be designed, Γ_T, and a random initialization of the energy matrix $\boldsymbol{\varepsilon}$, the scheme proceeds by iterating at each cycle the following steps:

(i) *Sequence selection*. Explore the combinatorial space of sequences to find the set $\mathcal{S} \equiv \{S_1, S_2, \ldots\}$ corresponding to the lowest values of $G(S)$.

(ii) *Structure prediction.* For each sequence in $\mathcal{S}$ obtain a reliable prediction of the native state. We shall indicate such a corresponding native set as $\mathcal{N} \equiv \{\tilde{\Gamma}_1, \tilde{\Gamma}_2, \ldots\}$.

(iii) *Energy function refinement*. Assess whether the states in $\mathcal{N}$ match the target structure Γ_T within a specified tolerance. If so, the design problem of Γ_T is solved, and the procedure ends. Otherwise, the symmetric energy matrix $\boldsymbol{\varepsilon}$ is refined to impose a consistency with the structure prediction results, i.e., to account for the fact that the native states of $\mathcal{S}$ do not include Γ_T.

A fixed point in this iterative scheme embodies the highest achievable consistency between our heuristic scoring function and the chosen protein structure prediction algorithm.

Step 1: Sequence selection. The first step of our iterative scheme involves solving a combinatorial optimization problem over the space of amino acid sequences. We will perform this step by constraining the overall amino acid composition, i.e., the abundances of the different types of amino acids. Thus, the first step involves minimizing Eq. (6) over the possible reshuffling of a given initial sequence with the desired composition. This task is an integer programming problem that can be carried out on conventional computers. However, considerable speedups for the same NP-complete class of problems may be achieved with quantum annealers, dedicated machines for solving quadratic unconstrained binary optimizations (QUBO) [64–67].

To recast the minimization of $G(S)$ in Eq. (6) as a QUBO problem, we introduced binary variables to describe the chemical type, s, of each amino acid in the sequence. Specif-

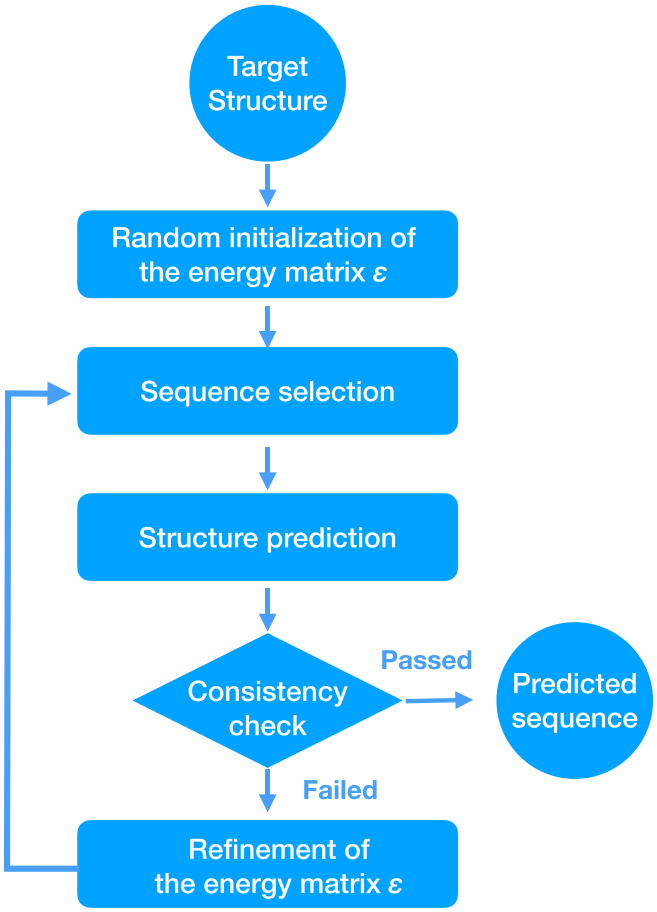

FIG. 1. Schematic representation of our protein-design algorithm. Alternating steps of sequence selection and structure prediction are repeated. The energy matrix from an initial guess is updated until the algorithm passes consistency checks. The result is a sequence that folds into the desired target structure. Importantly, sequence selection and structure prediction can be integrated from different algorithmic or hardware platforms, permitting leveraging on rapid and complementary progress in technologies such as machine learning and quantum computing.

ically, to the ith amino acid we associate an array of D binary variables, $q_m^{(i)}$, with $m = \{1, \ldots, D\}$, with the proviso that, if s_i is of type j, then $q_j^{(i)} = 1$ while the other $D-1$ array elements are equal to 0. Because of this, it suffices to directly specify the binary variables $q_2^{(i)}, \ldots, q_D^{(i)}$.

The variables are used to define a QUBO Hamiltonian whose ground states are in one-to-one correspondence with the sequences that minimize $G(S)$ at fixed composition. This QUBO Hamiltonian consists of three terms, all quadratic in the q entries,

$$\mathcal{H} = \mathcal{H}_{\rm comp} + \mathcal{H}_{\rm occ} + \mathcal{H}_{\rm contact}. \tag{7}$$

The first term attains its global minimum on sequences satisfying the composition constraint:

$$H_{\rm comp} = A_1 \sum_{m=2}^{D} \left(\sum_{i=1}^{n} q_m^{(i)} - N_m \right)^2, \tag{8}$$

where $A_1 > 0$, and N_m is the assigned number of amino acids of type m. Minimizing this term ensures that the nonzero entries of the q arrays are consistent with the assigned composition.

The second term penalizes cases in which more than one entry of the $q_m^{(i)}$ array is equal to 1:

$$H_{\rm occ} = A_2 \sum_{i=1}^{n} \sum_{m\neq n=2}^{D} q_m^{(i)} q_n^{(i)}, \tag{9}$$

with $A_2 > 0$. Minimizing this term along with $H_{\rm comp}$ ensures that the q arrays encode a well-defined chemical species for each amino acid in S.

The last term embodies the scoring function of Eq. (6):

$$H_{\rm contact} = B \Bigg[\sum_{i,j}{}' \sum_{m,n=2}^{D} q_m^{(i)} q_n^{(j)} C_{ij} \alpha_{mn} + 2 \sum_{ij}{}' \sum_{m=2}^{D} q_m^{(i)} C_{ij} \gamma_m + \sum_{i,j}{}' C_{ij} \varepsilon_{11} \Bigg]. \tag{10}$$

In this equation, $B > 0$ and $\boldsymbol{C} = \boldsymbol{C}(\Gamma_T) - \langle \boldsymbol{C} \rangle$, where $\boldsymbol{C}(\Gamma_T)$ is the adjacency matrix of the target configuration, and $\langle \boldsymbol{C} \rangle$ is the average contact map evaluated on a database of representative native structure. In addition, $\alpha_{mn} = \varepsilon_{mn} - \varepsilon_{m1} - \varepsilon_{n1} + \varepsilon_{11}$ and $\gamma_m = \varepsilon_{m1} - \varepsilon_{11}$.

The detailed derivation of this term is given in Sec. S1 of the Supplemental Material (SM) [68]. We note that the QUBO encoding of the minimization problem of Eq. (6) is not unique; an example of an alternative QUBO encoding is presented in Sec. S2 of the SM [68].

We emphasize that $H_{\rm occ}$ and $H_{\rm comp}$ encode the strong constraints, while $H_{\rm contact}$ represents the molecular energy. As long as $A_1, A_2 \gg B$, the ground-state solutions of Eq. (7) simultaneously satisfy all the hard constraints and correspond to sequences that minimize $G(s)$ for the given chemical composition.

The sought-after ground states of the QUBO Hamiltonian can be found with various methods, some of which are compared here, including classical simulated annealing (see Sec. S3 in the SM [68] for details), and the hybrid classical-quantum annealing scheme implemented in OCEAN, the user interface to the D-wave quantum annealer. The latter combines classical taboo search heuristic optimization with quantum annealing steps [69].

Step 2: Structure prediction. The second step of the iterative scheme involves the application of structure prediction methods to the sequences identified from the minimization of $G(S)$ at the previous step. The key point is that the native states of such sequences are obtained with an independent structure prediction method. In particular, the scheme used to predict native structures given a sequence is not informed by the energy matrix defining the design scoring function $G(S)$.

In realistic off-lattice applications, the go-to structure prediction methods would naturally be those based on heuristic machine-learning algorithms, which have proved to be reliable and efficient. Considering the protein model's minimalistic nature, we opted for the most transparent and feasible method: an exhaustive search of conformational space to identify the lowest energy state(s) of a sequence based on a

ground-truth energy matrix. This energy matrix is used solely in this step and for selecting viable target structures for the design problem, as detailed later, and is never used in the $G(S)$ definition.

Step 3: Energy function refinement. The third step in the proposed iterative scheme involves updating the energy matrix entering $G(S)$ to improve consistency with the chosen ground-truth structure prediction algorithm.

To this end, we have devised the following scheme: At the kth step of the iterative procedure, let S be a putative designing sequence obtained by minimizing G based on the current energy matrix, $\boldsymbol{\varepsilon}^{(k)}$. The external protein structure prediction algorithm may find several structures for the sequence S that fare better than Γ_T as native states. Let $\{\Gamma_0, \Gamma_1, \ldots, \Gamma_n\}$ be a ranked set of $n+1$ of such competing structures, ordered by increasing ground-truth energy, i.e., decreasing confidence score.

Since the structure prediction step is assumed to be reliable (and it certainly is in our minimalistic context where it entails an exhaustive search in structure space), the observation that the competing structures $\Gamma_0, \ldots, \Gamma_n$ have a higher confidence score than Γ_T signals the imperfect parametrization of the $\boldsymbol{\varepsilon}^{(k)}$ matrix.

To compensate for this, we move to a new iteration where the energy matrix $\boldsymbol{\varepsilon}^{(k+1)}$ is updated over the kth one by requiring that $\Gamma_0, \ldots, \Gamma_n$ have a lower energy than Γ_T, consistent with the outcome of the ground-truth predictor:

$$E^{(k+1)}(\Gamma_i, S) \leqslant E^{(k+1)}(\Gamma_T, S), \quad i \leqslant n, \tag{11}$$

where $E^{(k+1)}$ is the energy function of Eq. (4) informed by the interaction matrix $\boldsymbol{\varepsilon}^{(k+1)}$. Let us consider the case in which the best ranking structure Γ_0 has predicted probability above a given p_{fold}. In this case, we impose that the energy of Γ_0 should be significantly lower than that of all competing structures: $\forall i > 0$,

$$E^{(k+1)}(S, \Gamma_i) \geqslant E^{(k+1)}(S, \Gamma_0) + \Delta(p_{\text{fold}}, \beta), \tag{12}$$

where $\Delta(p_{\text{fold}}, \beta) = \frac{1}{\beta} \ln \frac{1-p_{\text{fold}}}{p_{\text{fold}}}$ is the minimum energy gap that would allow a sequence to fold into Γ_0 with probability $\geqslant p_{\text{fold}}$ at inverse temperature β (see Sec. S3 of the SM [68] for details).

Given that the energy coefficients to be learned enter linearly in the scoring function of Eq. (6), fulfilling the set of inequalities is equivalent to solving a linear separability problem. This task can be conveniently tackled using established algorithms [70–73]. In particular, in Sec. S4 of the SM [68], we discuss our implementation based on the perceptron technique [70,74–76], which has been previously used in different protein folding and design contexts to iteratively learn interaction potentials between amino acids [57,77].

C. Lattice protein model

We considered compact structures on a two-dimensional (square) lattice as a specific protein lattice model. The latter is customarily preferred over the cubic lattice because it offers a more realistic surface-to-volume ratio of compact structures of small length, $\lesssim$100 amino acids. We consider sequence alphabets of $D = 3$, 4, and 5 letters and target structures filling 4×4, 5×5, and 6×6 lattices.

The $D \times D$ symmetric energy matrix, embodying the ground-truth interaction potentials of the structure prediction step, was identified from a preliminary survey of viable combinations of interactions, meaning interactions that yield numerous designable structures. The latter correspond to structures that are the unique ground states of one or more sequences [10]. The choice of ground-truth energy matrix is detailed in Sec. S6 of the SM [68].

III. RESULTS

In our iterative design strategy, the optimal parameters of the scoring function G are obtained by comparing the results of direct versus inverse folding predictions. Several factors may determine the quality of the predictions: (i) the feasibility of minimizing G in the combinatorial space of sequences, (ii) the accuracy of the "external" structure prediction method, (iii) the viability of the functional form of G for yielding accurate design predictions when suitably parametrized, and (iv) the feasibility of identifying such optimal parametrizations of G using the iterative scheme.

In this proof-of-concept study, the uncertainties associated with points (ii) and (iii) are ruled out from the outset. Indeed, modeling proteins as compact structures on square lattices makes it possible to perform exhaustive searches in structure space, thus enabling the exact determination of the lowest energy state(s) of any given sequence. In addition, the functional form of the scoring function G, namely a pairwise-contacts Hamiltonian, was purposely chosen to match that of the ground-truth Hamiltonian used to pick designable structures as viable targets, thus guaranteeing that suitable parametrizations of G exist and are, in principle, learnable.

In our context, where we shall use alphabets of limited size, $D = 3, 4, 5$, point (i) could be addressed by exhaustive enumeration, similarly to the structure prediction step. However, in realistic contexts the exhaustive search of sequence or structure spaces would be unfeasible. While the structure space search can today be circumvented using the now available rapid and accurate structure prediction methods based on machine learning, the challenge of minimizing $G(S)$ in sequence space still persists. For this reason, we address point (i) by recasting the minimization of $G(S)$ as a QUBO problem, which can be tackled with classical and quantum combinatorial algorithms; see Sec. II. As we demonstrate by considering various values of D, such an approach is straightforwardly adapted to amino acid alphabets of any size.

In connection to points (ii) and (iii) outlined above, we first assess the viability of the approximate functional form of $G(S)$ as a scoring function to tackle the sequence optimization step. We considered two different contexts. In the first, the $\boldsymbol{\varepsilon}$ matrix is not learned but is set equal to that used in the ground-truth protein folding predictor. In the second, $\boldsymbol{\varepsilon}$ is learned through our iterative procedure.

To carry out this assessment, we take the compact structure Γ_T of Fig. 2(a) as the design target. We use an alphabet of $D = 3$ letters and set the composition to $[n_0 = 5, n_1 = 5, n_2 = 6]$, a choice that combines a sizable combinatorial space of sequences with the existence of numerous solutions to the design problem.

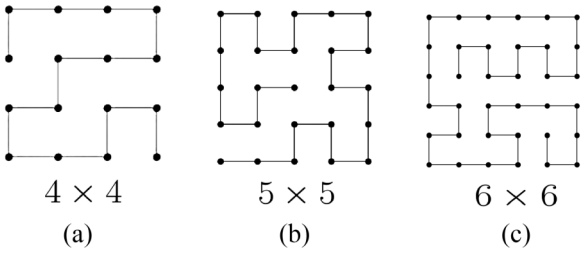

FIG. 2. Selected target structures for different lattice sizes.

For both $\boldsymbol{\varepsilon}$ choices, we computed $G(S)$ for all sequences with the above composition. We then obtained the receiver-operating curve (ROC), $y(x)$, where x is the rank index for increasing $G(S)$, and y indicates what fraction of the exhaustive set of design solutions are found up to that value of the scoring function.

A perfect design performance would yield the steepest ramping ROC curve, where all the design solutions are exclusively at the highest-ranking positions (lowest values of the scoring function). Accordingly, a customary measure of ROC performance is the so-called normalized area under the curve, Q, i.e., the area between the curve and the diagonal divided by the area of the upper triangle. The aforementioned perfect performance would correspond to $Q \sim 1$. In contrast, in a baseline performance—where solutions are discovered with uniform probability independent of their $G(S)$ ranking—Q would be close to 0.

The results of our ROC analysis are shown in Fig. 3. We preliminarily tested our $G(S)$ approximation by plugging the ground-truth potentials in place of the $\boldsymbol{\varepsilon}$ matrix. The corresponding ROC curve, shown with a dashed line in Fig. 3, shows a near-ideal performance, $Q > 0.99$. This demonstrates that the heuristic scoring function $G(S)$ of Eq. (6), which is based on average contact probabilities, is indeed viable for design purposes, as it can lead to a near-perfect scoring when informed by suitable potentials.

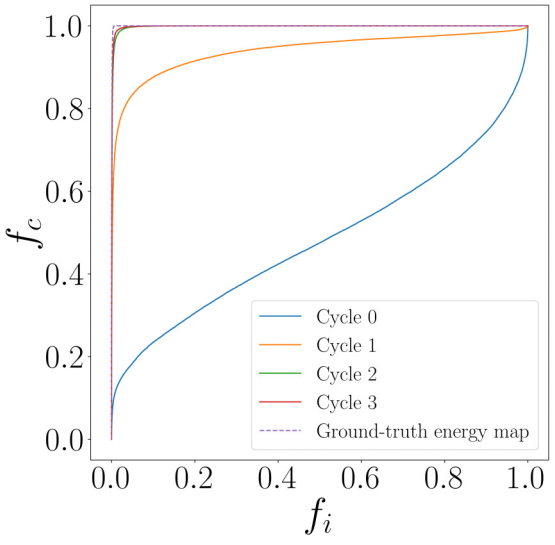

FIG. 3. Solid lines: ROC curve at different iterations tests the goodness of $G(S)$ [see Eq. (6)] as a classifier, showing that it evolves from being nearly random (at cycle 0) to nearly optimal (at cycle 3). Dashed line: Plugging the ground-truth energy map in $G(S)$ produces a nearly optimal classifier.

We then moved to the second assessment, aimed at ascertaining if suitable parametrizations of $G(S)$ can be learned by our iterative design procedure starting from arbitrary initializations of the $\boldsymbol{\varepsilon}$ matrix. A further question is how many iterations are required for convergence.

For these tests, we applied the iterative procedure to the same designable target structure starting from 50 different random choices of the initial matrix. The results are summarized in Fig. 3, which shows the ROC curves at different iteration stages, averaged over the different initializations—see Sec. S7 of the SM [68] for the individual ROC curves.

The blue curve in Fig. 3 shows the average performance at the beginning of the iterative procedure (labeled cycle 0) when the energy matrix is yet to be learned. The curve is near-diagonal in this case, demonstrating the expected baseline performance. The performance steadily improves at each iteration, converging to $Q > 0.99$, in as few as three iterations.

A. Performance scaling with lattice and alphabet size

We next turned to larger lattice sizes and amino acids alphabets; see the SM [68]. In such cases, ROC curves are not the best way to assess the design performance as they require exhaustive coverage of sequence space, which becomes rapidly impractical with growing protein length and alphabet size.

Instead, we estimated the design success rate using a sampling scheme. Specifically, at each iteration, we selected the 30 best-scoring sequences according to $G(S)$ and computed which fraction of them, f_c, admitted the target structure as the unique ground state and satisfied Eq. (2).

The results are given in Fig. 4(a) and show that, for all three alphabet sizes considered, $D = \{3, 4, 5\}$, our algorithm reaches a success rate of about 80% after just a few iterations. Notably, the highest performance is achieved with the largest alphabet size, corresponding to five amino-amino acid types. Importantly, this trend is robust over different choices of the target structure (see Sec. S8 of the SM [68]).

In Fig. 4(b), we report the results of a similar analysis for a fixed alphabet of $D = 3$ letters but for three compact structures filling lattices of increasing sizes. Again, in all cases, the algorithm reaches a plateau after a few iterations. While these specific instances do not show a clearly identifiable trend with lattice size, when the analysis is extended to an ensemble of structures, we observe that the success rate decreases with increasing chain size (see Sec. S8 of the SM [68]).

For the systems we have considered, the overall success rate ranges from about 65% up to nearly 100%.

B. Comparison of conventional and QUBO-based minimizers

A key feature of our approach is that the combinatorial search underpinning the sequence selection step is formulated as a QUBO problem, which is, in principle, amenable to quantum annealers. This poses two questions: (i) Does the QUBO encoding boost the design performance compared to working directly in sequence space? (ii) How do currently

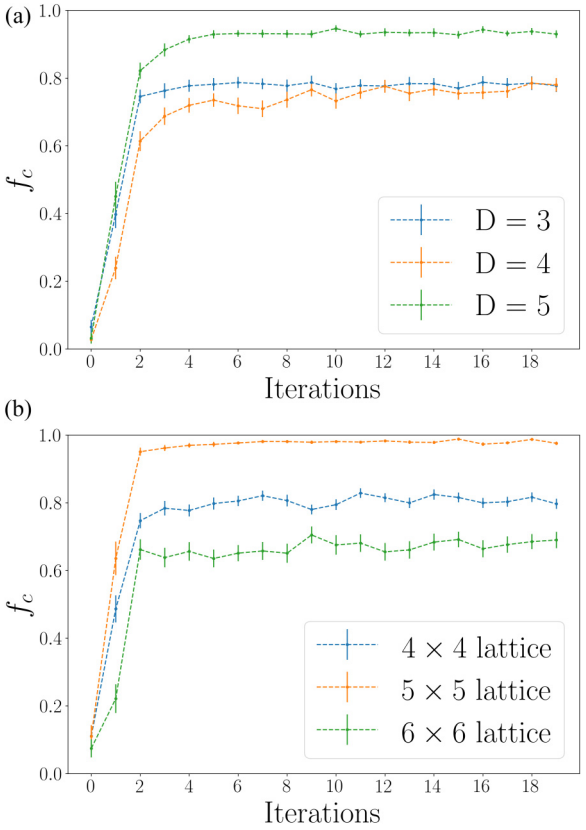

FIG. 4. Fraction f_c of correctly identified sequences as a function of refinement iterations for different alphabet and lattice sizes. In (a), we consider a fixed structure on a 4×4 lattice, see Fig. 2(a), and we vary the size D of the alphabet. In (b), we fix the alphabet size $D = 3$ and span over different lattice sizes. The corresponding target structures are in Figs. 2(a), 2(b), and 2(c).

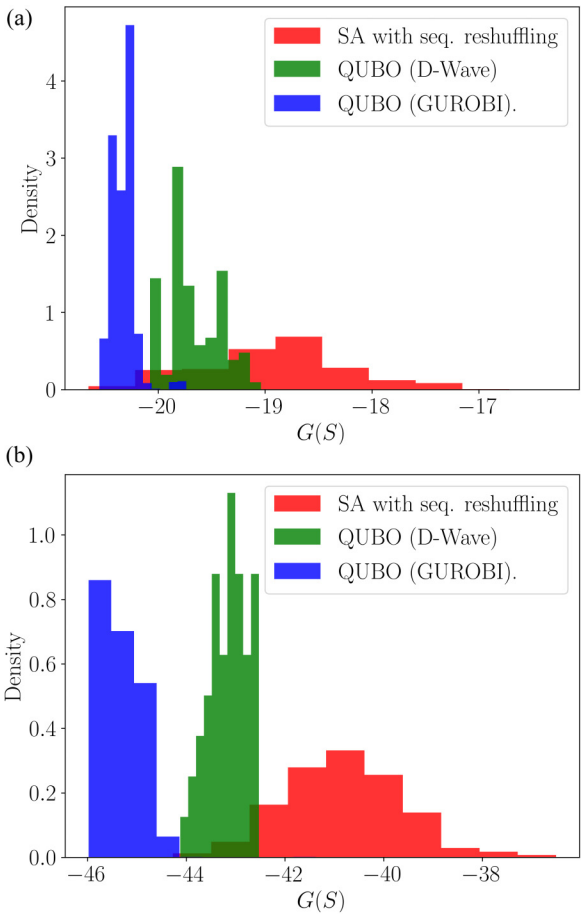

FIG. 5. Statistics of the $G(S)$ values resulting from the use of different optimization approaches (simulated annealing, hybrid annealing on D-WAVE, and GUROBI optimizer). In particular, we represent 1000 samples obtained by using simulated annealing, 1000 using the GUROBI optimizer, and 100 using the D-Wave hybrid optimizer. In (a) we consider a structure on a 9×9 lattice, while in (b) we consider a 13×13 lattice.

available quantum annealers fare at the design task compared to state-of-the-art classical QUBO solvers?

To address these questions, in Fig. 5 we report, for compact structures of different sizes, the lowest values of $G(S)$ obtained after many runs of classical optimizations of the scoring functions, parametrized with the ground-truth potentials, with different encodings and hardware at equal duration (3 s). The red histogram corresponds to the results of simulated annealing directly formulated in sequence space, where the moves correspond to composition-preserving reshufflings of the sequence. Instead, the blue curve was obtained using GUROBI, an industry-grade QUBO solver. Finally, the green histogram reports the results of 3 s runs on D-Wave using the hybrid classical-quantum solver, given that the complexity of the problem at hand exceeds the size currently addressable with fully quantum annealing algorithms.

The most striking feature of these results is that the $G(S)$ distributions of the minimizers based on the QUBO formulation (green and blue) extend well below the lower tail of the distribution generated with a conventional optimization based on the combination of sequence reshuffling and simulated annealing (red). Hence, the QUBO reformulation required to harness quantum computing technologies has generated a major improvement in the sequence optimization step even when adopted on classical machines. Notably, this difference in performance is enhanced for the largest lattice size.

Focusing on QUBO solvers, we note that the best performance is achieved by GUROBI, an entirely classical scheme based on heuristic searches. This result highlights the maturity reached by classical optimizers following from decades of hardware and software development. At the same time, we emphasize that the hybrid scheme implemented in D-Wave interleaves classical and quantum steps with internal criteria that are not easily controllable by the user. Thus, the results of the hybrid algorithm arguably represent a lower bound of the performance achievable by optimal combinations of classical and quantum steps.

IV. PATHWAY TOWARDS REALISTIC APPLICATIONS

In this section, we outline one possible route for transferring the method discussed so far to realistic off-lattice design contexts, and we highlight the challenges that still need to be overcome.

With reference to the flowchart of Fig. 1, this endeavor would involve three main directions: (i) the generalization of the design scoring function of Eq. (6) to off-lattice models, (ii) the use of tools such as AlphaFold for the structure prediction and energy refinement steps, and (iii) the generalization of the energy and free-energy terms of Eqs. (4) and (5) to account for the articulated amino acid structure, including rotameric states, and cooperative interactions. In addition to these conceptual points, one should additionally consider how the extension from a minimalistic to a realistic setting can impact the required computational resources.

Generalization to an off-lattice protein representation: Our protein design strategy is based on the scoring function of Eq. (6), where the structural information is encoded through contact maps. The discretized nature of lattice configurations is only reflected by the binary nature of the first matrix $\boldsymbol{C}$ in Eq. (6), i.e., the contact map of the target structure. However, the binary nature of C_{ij} is not a requirement of our method. This is manifest by the fact that the second matrix in Eq. (6) is an average contact map, and hence real-valued even in lattice contexts. In fact, the structural encoding based on contact maps is inherently robust because it does not hinge on explicit Cartesian coordinate representations nor the discreteness of the embedding space, and not even its dimensionality. Therefore, our formulation of the design scoring function is manifestly transferrable to off-lattice models.

Use of AlphaFold to perform protein structure predictions: The minimalistic lattice model used in our proof-of-concept application enabled us to obtain protein folding predictions from a predetermined (ground-truth) energy function, resorting to the exhaustive exploration of all compact protein structures.

The generalization from minimalistic to realistic models requires a reliable tool to perform protein structure predictions. AlphaFold provides the most accurate option for this task to date. In this scheme, the native structures are predicted by a deep neural network trained on a databank, not by a physics-based model. This structure prediction tool can be seamlessly integrated in our energy-refinement step. In fact, Alphafold calls do not return a single structure, but rather a set of ranked structure predictions. The ranking is based on a physical measure (pLDDT) that reflects the propensity of amino acids to adopt local structures different from the target one. Thus, the ranked set of structures and their pLDDt scores can be seamlessly used to select the competing structures to use in the energy-refinement step based on Eq. (11).

Improvement of the energy model in the scoring functional: Since the goal of the protein design problem is to identify sequences that yield a given *fold*, it is natural to resort to a scoring function based on a coarse-grained representation of the chain. Our minimalistic energy model is already equipped to account for pairwise interactions of the coarse-grained amino acids. However, in realistic applications, two-body interactions would not typically suffice, and effective many-body interactions would be needed to implicitly account for, e.g., internal degrees of freedom of the amino acids, such as rotameric states of the sidechains. This multibody approach is consistent with Anfisen's principle, which guarantees the possibility of defining a scoring function that discriminates between native and competing folds on the basis of the sole sequence information. Indeed, such expansions have been exploited before in accurate coarse-grained protein contexts, such as the UNited RESidue models developed by Scheraga and co-workers [78,79]. Thus, both the energy and free-energy expressions of Eqs. (4) and (5) may be extended through a many-body expansion,

$$E = E_1 + E_2 + E_3 + \cdots , \tag{13}$$

where E_k represents the k-body interaction.

As noted above, the minimalistic model discussed so far was based on retaining only the E_2 contribution. The inclusion in Eqs. (4) and (5) of unary terms, such as those needed to account for amino acid hydropathy or Ramachandran angle constraints, would be straightforward. Including three-body terms would additionally account for cooperative interactions. To this end, a natural form of three-body terms to include in Eqs. (4) and (5) would be

$$E_3 = \sum_{\mathrm{ijk}} \Gamma_{\mathrm{ijk}}(\Gamma)\, \epsilon_{q(\mathrm{i})q(\mathrm{j})q(\mathrm{k})}. \tag{14}$$

In this equation, $\Gamma_{\mathrm{ijk}}(\Gamma) \in [0, 1]$ is a rank-3 tensor determined entirely by the structure upon which a sequence $q(1), \ldots, q(N)$ is mounted. This tensor is defined to approach 1 only when residues i, j, and k are simultaneously within a cutoff distance. Similar expressions can be provided for the order-N terms and involve rank-N tensors.

An important challenge that remains to be tackled is how to efficiently incorporate explicit rotameric degrees of freedom, also required to correctly account for the side-chain and backbone packing.

Computational challenges for realistic protein design: Upgrading from a minimalistic to a realistic calculation will impact the computational cost of our algorithm. In particular, including many-body interactions and possibly rotameric states will greatly enlarge the space of parameters to be learned iteratively, thus requiring a larger training data set.

A possible strategy to balance the accuracy and the learning cost of the energy matrix could be to use physicochemical insight to restrict the structure of the many-body interactions, leaving only a few phenomenological parameters to be learned.

For example, one may harness three-body interactions to penalize the colocalization of triplets of amino acids that are seldom observed to be mutually proximal in protein data bank entries. This way, all such interactions would be weighted by the same coupling constant, i.e., an energy penalty. This practical and feasible scheme would greatly reduce the combinatorial search space due to the incorporation of phenomenological information, including steric constraints depending on rotameric degrees of freedom and cooperative amino acid interactions.

The inclusion of many-body terms would also affect the quantum encoding required to perform the sequence optimization step on a quantum annealing machine. Indeed, the current QUBO formulation is natively predisposed to deal with one- and two-body interactions. However, higher-order interactions can still be introduced by resorting to ancillary variables, as explicitly shown, e.g., in [46]. We expect that such demand of an increased number of qubits will be met by the rapid growth in size and performance of quantum

annealing machines. Our basic (two-body interactions only) QUBO formulation would require no more than about 1000 qubits for representing sequences of 100 amino acids with the full 20-letter chemical alphabet. For reference, the recent ground-breaking applications of DWAVE to a physics-based problem [80] have employed 5000 qubits.

Finally, regarding the computational cost of the structure prediction step, we note that a single call of AlphaFold-2 using the dedicated web server takes about 10 min to return the structure of a globular protein of about 100 amino acids. This time is shorter than that required in this work to perform the exhaustive search in the largest lattices considered, thus supporting the viability of integrating neural-network-based structure prediction tools.

V. CONCLUSION AND PERSPECTIVES

In this proof-of-concept study, we have shown that the availability of reliable algorithms for protein structure predictions can be capitalized to envision efficient strategies for tackling the protein design problem. Our iterative method has two distinctive features. First, the structure prediction algorithms are used to learn an optimal scoring function for the design problem instead of using postulated models and interaction parameters. Second, having mapped the sequence selection step to a combinatorial QUBO problem allows for addressing the design problem by harnessing existing powerful classical optimizers and promising quantum technologies.

Strikingly, we found that the QUBO encoding brings *per se* a significant improvement, to the point that matching the performance of classical or quantum QUBO optimizers with conventional schemes becomes computationally impractical even for modest protein lengths.

In our first illustrative benchmarks, we resorted to exhaustive enumeration to remove the uncertainty associated with heuristic protein folding predictors. This choice had the downside of considering simplified lattice models and relatively small chains and alphabets. However, we note that the key theoretical ingredient of our method is the scoring function $G(S)$ that is entirely specified in terms of contact maps, regardless of the specific representation of protein conformations. As such, this feature is key to envisioning generalizations to off-lattice contexts. However, extending our approach to include rotameric degrees of freedom is likely to be the necessary step to advance the method towards realistic applications.

State-of-the-art machine-learning predictors such as AlphaFold [33] provide the key to obtaining structure predictions with atomistic resolution. Notably, these schemes return accurate folding solutions in a time much shorter than what is required by our exhaustive enumeration in lattice models.

Furthermore, the trajectory in addressing the technological limitations of the existing quantum hardware has been impressive [64,81]. This gives hope that it will become feasible to introduce more accurate energy models, more elaborate structural representations for the amino acids, and to extend the sequence alphabet space.

Succeeding at these tasks would be transformative for protein design while paving the way towards a broad range of related applications, e.g., protein origami, drug screening, and *de novo* drug design.

ACKNOWLEDGMENTS

We are grateful to Francesco Slongo for his help with numerical work, and to Davide Pastorello and Enrico Blanzieri for useful discussions. This work was supported partly by Qub-IT, a project funded by the Italian Institute of Nuclear Physics (INFN) within the Technological and Interdisciplinary Research Commission (CSN5). C.M. acknowledges support from the European Union–NextGenerationEU, in the framework of the PRIN Project "The Physics of Chromosome Folding" (code: 2022R8YXMR, CUP: G53D23000820006) and by PNRR Mission 4, Component 2, Investment 1.4_CN_00000013_CN-HPC: National Centre for HPC, Big Data and Quantum Computing–spoke 7 (CUP: G93C22000600001). P.H. and V.P. acknowledge support from Q@TN, the joint laboratory of the University of Trento, FBK-Fondazione Bruno Kessler, INFN-National Institute for Nuclear Physics, and CNR-National Research Council. P.H. has further received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101080086 NeQST and the Italian Ministry of University and Research (MUR) through the FARE grant for the project DAVNE (Grant No. R20PEX7Y3A). This project was funded by the European Union under NextGenerationEU via the ICSC–Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing. P.H. also acknowledges support from Provincia Autonoma di Trento and from ICSC–Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing, funded by the European Union under NextGenerationEU. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union, The European Research Executive Agency, or the European Commission. Neither the European Union nor the granting authority can be held responsible. P.F. acknowledges support from by U. Milan-Bicocca's Center for Quantum Technologies (BiQuTe).

P.F. is a cofounder and shareholder of Sibylla Biotech SPA, a company developing and employing advanced molecular simulations for early-stage drug discovery.

[1] A. M. Lesk, *Introduction to Protein Science: Architecture, Function and Genomics* (Oxford University Press, Oxford, UK, 2004).

[2] K. A. Dill and H. S. Chan, From levinthal to pathways to funnels, Nat. Struct. Mol. Biol. **4**, 10 (1997).

[3] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, How fast-folding proteins fold, Science **334**, 517 (2011).

[4] C. B. Anfinsen, Principles that govern the folding of protein chains, Science **181**, 223 (1973).

[5] C. Anfinsen and H. A. Scheraga, Experimental and theoretical aspects of protein folding, Adv. Protein Chem. **29**, 205 (1975).

[6] C. Levinthal, Are there pathways for protein folding? J. Chim. Phys. **65**, 44 (1968).

[7] C. Levinthal, How to fold graciously, in *Mössbaun Spectroscopy in Biological Systems Proceedings* (University of Illinois Press, Urbana, Illinois, 1969), Vol. 67, pp. 22–26.

[8] J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), J. Phys. Chem. **93**, 6902 (1989).

[9] A. Šali, E. Shakhnovich, and M. Karplus, How does a protein fold? Nature (London) **369**, 248 (1994).

[10] H. Li, R. Helling, C. Tang, and N. Wingreen, Emergence of preferred structures in a simple model of protein folding, Science **273**, 666 (1996).

[11] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Chain length scaling of protein folding time, Phys. Rev. Lett. **77**, 5433 (1996).

[12] D. Klimov and D. Thirumalai, Cooperativity in protein folding: From lattice models with sidechains to real proteins, Folding Des. **3**, 127 (1998).

[13] H. S. Chan, S. Shimizu, and H. Kaya, Cooperativity principles in protein folding, in *Methods in Enzymology* (Elsevier, Amsterdam, 2004), Vol. 380, pp. 350–379.

[14] N. C. Fitzkee and G. D. Rose, Reassessing random-coil statistics in unfolded proteins, Proc. Natl. Acad. Sci. USA **101**, 12497 (2004).

[15] J. N. Onuchic and P. G. Wolynes, Theory of protein folding, Curr. Opin. Struct. Biol. **14**, 70 (2004).

[16] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, The protein folding problem, Annu. Rev. Biophys. **37**, 289 (2008).

[17] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[18] M. Negri, G. Tiana, and R. Zecchina, Native state of natural proteins optimizes local entropy, Phys. Rev. E **104**, 064117 (2021).

[19] J. M. Deutsch and T. Kurosky, New algorithm for protein design, Phys. Rev. Lett. **76**, 323 (1996).

[20] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Optimal protein design procedure, Phys. Rev. Lett. **77**, 1901 (1996).

[21] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Protein design in a lattice model of hydrophobic and polar amino acids, Phys. Rev. Lett. **80**, 2237 (1998).

[22] A. Irbäck, C. Peterson, F. Potthast, and E. Sandelin, Monte Carlo procedure for protein design, Phys. Rev. E **58**, R5249 (1998).

[23] E. I. Shakhnovich, Protein design: A perspective from simple tractable models, Folding Des. **3**, R45 (1998).

[24] P. Huang, S. E. Boyken, and D. Baker, The coming of age of de novo protein design, Nature (London) **537**, 320 (2016).

[25] I. V. Korendovych and W. F. DeGrado, *De novo* protein design, a retrospective, Q. Rev. Biophys. **53**, e3 (2020).

[26] A. Paladino, F. Marchetti, S. Rinaldi, and G. Colombo, Protein design: From computer models to artificial intelligence, WIREs Comput. Mol. Sci. **7**, e1318 (2017).

[27] C. Goverde, B. Wolf, H. Khakzad, S. Rosset, and B. E. Correia, De novo protein design by inversion of the AlphaFold structure prediction network, bioRxiv 2022.12.13.520346 (2022).

[28] M. Jendrusch, J. O. Korbel, and S. K. Sadiq, AlphaDesign: A de novo protein design framework based on AlphaFold, bioRxiv 2021.10.11.463937 (2021).

[29] T. Takahashi, G. Chikenji, and K. Tokita, Lattice protein design using Bayesian learning, Phys. Rev. E **104**, 014404 (2021).

[30] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, and D. Baker, De novo protein design by deep network hallucination, Nature (London) **600**, 547 (2021).

[31] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, De novo design of protein structure and function with RF diffusion, Nature (London) **620**, 1089 (2023).

[32] J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, and G. Grigoryan, Illuminating protein space with a programmable generative model, Nature (London) **623**, 1070 (2023).

[33] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).

[34] S. -J. Chen, M. Hassan, R. L. Jernigan, K. Jia, D. Kihara, A. Kloczkowski, S. Kotelnikov, D. Kozakov, J. Liang, A. Liwo, S. Matysiak, J. Meller, C. Micheletti, J. C. Mitchell, S. Mondal, R. Nussinov, K. Okazaki, D. Padhorny, J. Skolnick, T. R. Sosnick, G. Stan, I. Vakser, X. Zou, and G. D. Rose, Protein folds vs. protein folding: Differing questions, different challenges, Proc. Natl. Acad. Sci. USA **120**, e2214423119 (2023).

[35] A. Perdomo, C. Truncik, I. Tubert-Brohman, G. Rose, and A. Aspuru-Guzik, Construction of model hamiltonians for adiabatic quantum computation and its application to finding low-energy conformations of lattice protein models, Phys. Rev. A **78**, 012320 (2008).

[36] A. Perdomo-Ortiz, N. Dickson, M. Drew-Brook, G. Rose, and A. Aspuru-Guzik, Finding low-energy conformations of lattice protein models by quantum annealing, Sci. Rep. **2**, 571 (2012).

[37] A. Robert, P. K. Barkoutsos, S. Woerner, and I. Tavernelli, Resource-efficient quantum algorithm for protein folding, npj Quantum Inf. **7**, 38 (2021).

[38] M. H. Khatami, U. C. Mendes, N. Wiebe, and P. M. Kim, Gate-based quantum computing for protein design, PLoS Comput. Biol. **19**, e1011033 (2023).

[39] A. Irbäck, L. Knuthson, S. Mohanty, and C. Peterson, Folding lattice proteins with quantum annealing, Phys. Rev. Res. **4**, 043013 (2022).

[40] C. Micheletti, F. Seno, and A. Maritan, Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies, Proteins: Struct. Funct. Bioinf. **40**, 662 (2000).

[41] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, Small libraries of protein fragments model native protein structures accurately, J. Mol. Biol. **323**, 297 (2002).

[42] A.-C. Camproux, R. Gautier, and P. Tuffery, A hidden Markov model derived structural alphabet for proteins, J. Mol. Biol. **339**, 591 (2004).

[43] A. Pandini, A. Fornili, and J. Kleinjung, Structural alphabets derived from attractors in conformational space, BMC Bioinf. **11**, 97 (2010).

[44] C. O. Mackenzie, J. Zhou, and G. Grigoryan, Tertiary alphabet for the observable protein structural universe, Proc. Natl. Acad. Sci. USA **113**, E7438 (2016).

[45] P. Krupa, A. Hałabis, W. Zmudzinska, S. Ołdziej, H. A. Scheraga, and A. Liwo, Maximum likelihood calibration of the unres force field for simulation of protein structure and dynamics, J. Chem. Inf. Model. **57**, 2364 (2017).

[46] F. Slongo, P. Hauke, P. Faccioli, and C. Micheletti, Quantum-inspired encoding enhances stochastic sampling of soft matter systems, Sci. Adv. **9**, eadi0204 (2023).

[47] R. Babbush, P. J. Love, and A. Aspuru-Guzik, Adiabatic quantum simulation of quantum chemistry, Sci. Rep. **4**, 6603 (2014).

[48] R. Wong and W.-L. Chang, Fast quantum algorithm for protein structure prediction in hydrophobic-hydrophilic model, J. Parallel Distrib. Comput. **164**, 178 (2022).

[49] P. Hauke, G. Mattiotti, and P. Faccioli, Dominant reaction pathways by quantum computing, Phys. Rev. Lett. **126**, 028104 (2021).

[50] D. Ghamari, P. Hauke, R. Covino, and P. Faccioli, Sampling rare conformational transitions with a quantum computer, Sci. Rep. **12**, 16336 (2022).

[51] C. Micheletti, P. Hauke, and P. Faccioli, Polymer physics by quantum computing, Phys. Rev. Lett. **127**, 080501 (2021).

[52] K. F. Lau and K. A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules **22**, 3986 (1989).

[53] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, A test of lattice protein folding algorithms, Proc. Natl. Acad. Sci. USA **92**, 325 (1995).

[54] A. Irbäck, L. Knuthson, S. Mohanty, and C. Peterson, Using quantum annealing to design lattice proteins, Phys. Rev. Res. **6**, 013162 (2024).

[55] V. N. Maiorov and G. M. Crippen, Contact potential that recognizes the correct folding of globular proteins, J. Mol. Biol. **227**, 876 (1992).

[56] G. M. Crippen, Easily searched protein folding potentials, J. Mol. Biol. **260**, 467 (1996).

[57] F. Seno, C. Micheletti, A. Maritan, and J. R. Banavar, Variational approach to protein design and extraction of interaction potentials, Phys. Rev. Lett. **81**, 2172 (1998).

[58] A. Rossi, C. Micheletti, F. Seno, and A. Maritan, A self-consistent knowledge-based approach to protein design, Biophys. J. **80**, 480 (2001).

[59] R. L. Miyazawa and S. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation, Macromolecules **18**, 534 (1985).

[60] J. Skolnick, A. Godzik, L. Jaroszewski, and A. Kolinski, Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Prot. Sci. **6**, 676 (1997).

[61] G. Tiana, M. Colombo, D. Provasi, and R. A. Broglia, Deriving amino acid contact potentials from their frequencies of occurrence in proteins: A lattice model study, J. Phys.: Condens. Matter **16**, 2551 (2004).

[62] W. W. Chen and E. I. Shakhnovich, Lessons from the design of a novel atomic potential for protein folding, Prot. Sci. **14**, 1741 (2005).

[63] R. A. Goldstein, Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: Insights from the quasi-chemical approximation, Prot. Sci. **16**, 1887 (2007).

[64] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, Perspectives of quantum annealing: Methods and implementations, Rep. Prog. Phys. **83**, 054401 (2020).

[65] C. C. Chang, C.-C. Chen, C. Koerber, T. S. Humble, and J. Ostrowski, Integer programming from quantum annealing and open quantum systems, arXiv:2009.11970.

[66] J.-R. Jiang and C.-W. Chu, Solving np-hard problems with quantum annealing, in *Proceedings of the 2022 IEEE 4th Eurasia Conference on IOT, Communication and Engineering (ECICE)* (IEEE, Piscataway, NJ, 2022).

[67] R. Au-Yeung, N. Chancellor, and P. Halffmann, Np-hard but no longer hard to solve? using quantum computing to tackle optimization problems, Front. Quantum Sci. Technol. **2**, 1128576 (2023).

[68] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PRXLife.2.043012 for supplementary figures and discussion. We detail the derivation of the QUBO functional, the simulated annealing protocol, and the refinement procedure of the entries of the energy matrix. We include the numerical parameters adopted in our simulations and an analysis regarding the effect of increasing alphabet or lattice size on the overall success rate of our algorithm.

[69] J. Raymond, R. Stevanovic, W. Bernoudy, K. Boothby, C. C. McGeoch, A. J. Berkley, P. Farré, J. Pasvolsky, and A. D. King, Hybrid quantum annealing for larger-than-QPU lattice-structured problems, ACM Trans. Quantum Comput. **4**, 17 (2023).

[70] W. Krauth and M. Mezard, Learning algorithms with optimal stability in neural networks, J. Phys. A **20**, L745 (1987).

[71] D. G. Kleinbaum, *Logistic Regression. Statistics in the Health Sciences* (Springer, New York, NY, 1994), pp. 1–38.

[72] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).

[73] R. A. Johnson and A. Mouhab, A Bayesian decision theory approach to classification problems, J. Multivar. Anal. **56**, 232 (1996).

[74] J. Imhoff, A polynomial training algorithm for calculating perceptrons of optimal stability, J. Phys. A **28**, 2173 (1995).

[75] Y. Freund and R. E. Schapire, Large margin classification using the perceptron algorithm, Mach. Learn. **37**, 277 (1999).

[76] K.-L. Du, C.-S. Leung, W. H. Mow, and M. N. Swamy, Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era, Mathematics **10**, 4730 (2022).

[77] R. I. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan, Extraction of interaction potentials between amino acids from native protein structures, J. Chem. Phys. **112**, 9151 (2000).

[78] A. Liwo, J. Lee, D. R. Ripoll, and H. A. Scheraga, Protein structure prediction by global optimization of a potential energy function, Proc. Natl. Acad. Sci. USA **96**, 5482 (1999).

[79] Odziej*et al.*, Physics-based protein-structure prediction using a hierarchical protocol based on the unres force field: Assess-

ment in two blind tests. Proc. Natl. Acad. Sci. USA **102**, 7547 (2005).

[80] A. D. King *et al.*, Quantum critical dynamics in a 5,000-qubit programmable spin glass, Nature (London) **617**, 61 (2023).

[81] T. Pochart, P. Jacquot, and J. Mikael, On the challenges of using d-wave computers to sample boltzmann random variables, in *Proceedings of the 2022 IEEE 19th International Conference on Software Architecture Companion (ICSA-C)* (IEEE, Piscataway, NJ, 2022).