*Journal of Complex Networks* (2022) **1**, Advance Access Publication on 28 December 2021
https://doi.org/10.1093/comnet/cnab048

# Could network structures generated with simple rules imposed on a cubic lattice reproduce the structural descriptors of globular proteins?

OSMAN BURAK OKAN
*Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA and Coating Technologies Division, Sisecam Science and Technology Center, Kocaeli, Turkey*
DENIZ TURGUT
*Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA and Seven Bridges Genomics, Istanbul, Turkey*
CANAN ATILGAN AND ALI RANA ATILGAN
*Faculty of Natural Sciences and Engineering, Sabanci University, Istanbul, Turkey*

AND

RAHMI OZISIK[†]
*Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[†]Corresponding author. Email: ozisik@rpi.edu

Edited by: Ernesto Estrada

[Received on 2 July 2021; editorial decision on 1 October 2021; accepted on 30 November 2021]

A direct way to spot structural features that are universally shared among proteins is to find analogues from simpler condensed matter systems. In the current study, the feasibility of creating ensembles of artificial structures that can automatically reproduce a large number of geometrical and topological descriptors of globular proteins is investigated. Towards this aim, a simple cubic (SC) arrangement is shown to provide the best background lattice after a careful analysis of the residue packing trends from 210 globular proteins. It is shown that a minimalistic set of rules imposed on this lattice is sufficient to generate structures that can mimic real proteins. In the proposed method, 210 such structures are generated by randomly removing residues (beads) from clusters that have a SC lattice arrangement such that all the generated structures have single connected components. Two additional sets are prepared from the initial structures via random relaxation and a reverse Monte Carlo simulated annealing algorithm, which targets the average radial distribution function (RDF) of 210 globular proteins. The initial and relaxed structures are compared to real proteins via RDF, bond orientational order parameters and several descriptors of network topology. Based on these features, results indicate that the structures generated with 40% occupancy closely resemble real residue networks. The structure generation mechanism automatically produces networks that are in the same topological class as globular proteins and reproduce small-world characteristics of high clustering and small shortest path lengths. Most notably, the established correspondence rules out icosahedral order as a relevant structural feature for residue networks in contrast to other amorphous systems where it is an inherent characteristic. The close correspondence is also observed in the vibrational characteristics as computed from the Anisotropic Network Model, therefore hinting at a non-superficial link between the proteins and the defect laden cubic crystalline order.

*Keywords*: protein structure; residue network; coarse-graining; disordered systems; bond orientational order; elastic network models.

## 1. Introduction

The native protein structure is uniquely realized because it concurrently satisfies several constraints that are brought about by chain connectivity the feasibility of non-bonded interactions within the chain and with the solvent, excluded volume effects and side-chain placements [1, 2]. Despite this apparent complexity, major structural features of proteins and their functional implications can be studied by coarse-grained models [3–9]. Coarse-graining methods have been originally developed to simplify the structure, thereby increasing the attainable length and time scales of computations. Although the underlying principles of coarse-graining are similar to those used in other materials, that is, polymers, coarse-graining of proteins is particularly targeted to describe the native state. This has led to the development of various coarse-graining techniques such as Go models [4, 10–13], whose only objective is to fold a given protein assuming a priori knowledge of native topology.

In the current study, we aim to generate coarse-grained model structures comprising assemblies of hard spheres that mimic the geometrical and topological characteristics of real proteins with no reference to specific protein structures. Towards this target, we adopt a construction method with a small number of simple and straightforward rules imposed on a simple cubic (SC) lattice. The ground rules for structure generation are motivated by the average residue packing trends of 210 globular proteins.

The packing characteristics of geometrical objects offer a simple, yet extremely general, paradigm to study condensed matter systems [14–19]. Packing approaches have found several uses in protein research. At the most fundamental level, the hydrophobic core's efficient burial is shown to be a central requirement for protein folding [20–22]. Packing density is responsible for the distinct features on the potential energy landscape, and even a simple folded polypeptide chain has an inherently complex and multidimensional landscape, whereas that of a fully extended conformation is featureless [22, 23]. Tessellation studies showed that residue centres' space-filling characteristics carry signatures of Bernal packing of hard spheres [24], and free volume distribution resembles that of simple dense liquids [25]. At the tertiary structure level, the optimal size of protein domains can be deduced with simple sphere models and efficient burial of the hydrophobic core [26]. Finally, proteins have remarkably low intrinsic compressibility, yet they are amenable to unfolding under hydrostatic pressure [27–30]. The unfolding pressure is strongly dependent on the free volume distribution, and unfolding is accompanied by a reduction in specific volume [29]. For globular proteins, the presence of interior voids is now well documented [31–33], and more recent analyses of packing hint that the protein core is more open than previously reported [34, 35].

In the current study, we investigate the feasibility of using primitive cubic order combined with three sources of disorder: surface/finite-size effects, voids and positional deviations from ideal lattice sites. It is known that the distribution of internal coordinates of $C_\alpha$ atoms peaks around preferential values, which are compatible with cubic lattice geometry [36], and crystal lattices are frequently deployed in protein modelling efforts. The underlying lattice provides a basic grid for realizing and updating conformations [10, 36–38]. There are many folding algorithms based on these ideas [39] that make use of various lattice types. Notably, Covell and Jernigan used a face-centred cubic (FCC) lattice to generate all possible conformations and chose the optimal conformation based on non-bonded pairwise potential energy minimization [40]. Similarly, a lattice model based on the diamond cubic lattice (equivalent to an FCC lattice with a two-point basis) was introduced to predict folded conformations at low spatial resolution with no reference to a native state [41]. Simple cubic (SC) geometry is the simplest crystalline alternative and it has been used to study sequence–structure relationships at the residue level through exhaustive state enumeration and designability ideas based on the number of sequences that can realize a specific conformation [42]. Thus, it has been possible to distinguish between structures that are realized by multitude of sequences and are highly stable against mutations from compact structures with low designability due to strict sequence dependency. More recent approaches that utilize SC lattice incorporate

sequence specificity in additional detail and successfully probe folding events, peptide aggregation and larger scale phenomena such as amyloid formation. However, in these works, the underlying lattice is mainly a book-keeping device that restricts the conformational space [43–45].

In this study, we present a novel method for structure ensemble generation which is based on SC lattice clusters where the structural constraints imposed by this coordination geometry is a fundamental ingredient in structure generation. This lattice choice is motivated by an attempt to directly map protein packing trends to a suitable lattice template. The presented hard sphere model for structure generation is novel in the sense that it enables us to generate an ensemble of structures at once where each generated member is representative of the geometrical and topological trends observed in globular proteins. We provide evidence that the broad correspondence between model structures and real proteins can only be realized with a SC lattice background. To generate the desired model structures on a SC lattice, an extra randomly discarded percentage is needed to capture the packing trends observed from the experimental coordinates of $C_\alpha$ atoms.

The comparisons rely on the use of experimental coordinates and no attempt is made to elucidate the protein folding mechanisms or related dynamics. Excessive introduction of unoccupied sites to the core regions of the generated structures is avoided by two constraints: the single-component connectivity of the retained sites and the high surface to volume ratio of the starting clusters. A simple explanation of the degeneracy and the ruggedness of the potential energy landscape is attempted with these basic tools. Our approach is in a way similar to the Watts–Strogatz approach, wherein one generates small-world networks starting from a uniform circular graph and reaches graphs of various clustering, and average path lengths, controlled by the degree of rewiring introduced [46]. It is indeed shown that the universal topological class typically populated by proteins but inaccessible to random network growth mechanisms [47] can be recovered with our approach. The resulting structures automatically show small-world characteristics such as high local clustering and small shortest path lengths [46].

To probe the similarities between the generated structures and real proteins, we investigate the distributions of fundamental geometrical and topological descriptors. Through ensemble-based structural comparisons, we are able to quantify the extent of non-superficial similarities between the defect laden cubic order and residue networks. Geometrical comparisons are based on the local bond orientational order (BOO) parameters, and the topological comparisons are based on several network metrics that are directly defined from the moments of the adjacency matrix.

## 2. Computational methods

The generation of model protein structures is performed using three simple rules: the presence of a molecular surface, the inclusion of free space (voids) and positional deviations from ideal lattice sites. These rules are applied in the order they are listed, and the details of the computational methods used in the generation of the artificial structures are provided in Section 2.1. In Section 2.2, we outline the structural characterization methods used to compare the generated structures to the average protein structure.

### 2.1 *Generation of artificial structures*

2.1.1 *Formation of finite-sized lattice clusters* In the current study, each lattice point is considered to represent an amino acid (residue) located at the $C_\alpha$ atom. Initially, a finite-sized cluster is carved out of the bulk SC crystal with a lattice spacing of 3.8 Å. Throughout this study, 210 different such clusters are randomly generated. These 210 structures are equally selected (35 of each) from 6 different sized clusters: $5 \times 5 \times 6, 6 \times 6 \times 7, 7 \times 7 \times 7, 7 \times 7 \times 8, 8 \times 8 \times 9$ and $8 \times 9 \times 9$ to capture the size variations

in the comparison set of 210 proteins. The shape of the SC clusters is chosen to be approximately cubic because it is the most convenient choice; it is compatible with the basic structure of the underlying SC lattice, and it extends equally along with the basal directions of the SC lattice. However, it is important to note that clusters of differing geometries were also tried, and that the results do not depend on the shape of the cluster as long as elongated shapes with very high aspect ratios ($>1.8$) are not used was verified.

2.1.2 *Inclusion of empty space* Randomly selected residues (coarse-grained beads) are removed from a finite-sized cluster until a pre-determined void concentration is reached. During the introduction of voids, the structure is checked for connectivity after the removal of each residue. Connectivity is defined such that at any stage during the introduction of voids, a path exists between any two randomly selected residues. Since there are no universal rules regarding the spatial distribution of voids, both internal and surface voids are generated without any constraints. The resulting structures are three-dimensional bead clusters each having a single connected component; these are not necessarily self-avoiding walks because of the way the connectivity is defined during the removal process.

2.1.3 *Introduction of positional disorder* The structures generated in the previous section are made up of coarse-grained beads located at the lattice points. However, RDFs pertaining to such clusters comprise successions of delta functions and unlike the case for real proteins they do not show any spread. Therefore, it is necessary to relax the residues of the generated structures from their on-lattice positions. This relaxation (the introduction of positional disorder) is achieved through the displacement of coarse-grained beads (residues) via two alternative methods. The first method moves individual beads by targeting the average RDF of real proteins by a reverse Monte Carlo and simulated annealing (RMC–SA) algorithm. The second method invokes displacements by randomly moving residues. The latter is terminated once the average displacement of the beads reaches 0.5 Å. This particular choice ensures that the average displacement is much smaller than the nearest neighbour spacing and is adopted in light of average displacements attained by the RMC–SA method described below.

The RMC procedure is a reconstruction method that aims to evolve a structure until the selected property (or properties) of the evolving structure mimics or best represents the selected experimental property [48, 49]. SA is a heuristic global optimization method where the probability of accepting solutions is continuously decreased as the solution space is explored. In the current study, the target property of the RMC–SA procedure is chosen to be the average RDF of the 210 basis proteins up to 15 Å distance.

The RMC optimization proceeds with Boltzmann sampling using a simple Metropolis scheme [50]. Each attempted move is accepted with a probability ($p$), see Eq. (1), where $\Delta E$ is the energy difference between the attempted and current configurations of the relaxing structure. The energy function, $E$, is defined as the integrated squared difference between the RDF of the generated structure, $g(r_{i,t})$, and the average RDF of the 210 basis proteins, $\bar{g}$(r), up to 15 Å, see Eq. (2).

$$P = \min(1, e^{\beta \Delta E}) \tag{1}$$

$$E(r_{i,t}) = \int_0^{15} \left[ g(r_{i,t}) - \bar{g}(r) \right]^2 dr. \tag{2}$$

In Eq. (2), the index $i$ represents current or attempted structures at step $t$, and the Boltzmann factor is defined as follows:

$$\beta^{-1} = \alpha E(r_{\text{current},t=0}), \tag{3}$$

where $\alpha = 10^{-6}$ is used as a scaling constant. The simulated annealing is implemented by updating $\beta^{-1}$ by 0.9 at every $10^3$ steps. The combined RMC–SA procedure is performed for $5 \times 10^5$ steps or until $\beta\Delta E \leq 10^{-7}$ for at least $2 \times 10^3$ consecutive steps.

### 2.2 *Characterization of structural features*

The comparison of the generated structures with an average protein as defined from the 210 globular proteins is performed based on three main features: residue packing, residue bonding anisotropy and topology (connectivity of residues). The packing consideration is based on the radial distribution function and nearest neighbour considerations. The bonding anisotropy is investigated via the bond orientational order (BOO) parameters. Finally, the contact topology is investigated using network analysis tools. In the next sections, we briefly describe BOO, network parameters and the elastic network model construction used in the current study.

2.2.1 *Bond orientational order* BOO quantifies the anisotropy of bonding; its parameters are defined as the rotationally invariant combinations of bond density expansion coefficients. The bond density expansion is carried out with spherical harmonics which form a basis for the $(2l+1)$-dimensional representation of the rotation group SO(3), and therefore, they are well suited to study three-dimensional systems with finite symmetries whose rotational symmetries conform to finite subgroups of the rotational group [51]. In the protein community, BOO has been used to define anisotropic statistical potentials, to study the stability of collective motions, and to assign secondary structures [6, 52–54].

Originally BOO [55] is defined as a global order parameter, but we will use the local definition that accounts for the bond vectors in the coordination shell for each constituent residue:

$$Q_{l,i} = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{Q}_{l,m,i}|^2}. \tag{4}$$

Here, $\bar{Q}_{l,m,i}$ denotes average over all bonds emanating from a residue $i$ and is defined as follows:

$$\bar{Q}_{l,m,i} = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{l,m}(\theta_{i,j}\varphi_{i,j}), \tag{5}$$

where $Y_{l,m}\left(\theta_{ij}, \phi_{ij}\right)$ are spherical harmonics[1] computed at the polar, $\theta_{ij}$, and azimuthal, $\varphi_{ij}$, angular coordinates of the unit vector $\mathbf{r}_{ij}$ directed from residue $i$ to its neighbour $j$. Note that the final expression of the order parameter $Q_{l,i}$ in Eq. (4) does not depend on $m$ for rotational invariance.

It is possible to systematically define higher-order invariants with spherical harmonics [55]. The simplest and physically the most interesting higher-order invariants are the properly normalized third

[1] For spherical harmonics, we adopted the standard definition and ignored the Condon–Shortley phase $(-1)^m$, which leads to $Y_{l,m}(\theta,\varphi) = \left[\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}\right]^{1/2} P_{l,m}(\cos(\theta))e^{im\varphi}$, where $P_{l,m}(\cos(\theta))$ are the associated Legendre functions. Normalization constant is consistent with Steinhardt *et al.* [50] and yields $Y_{0,0} = \frac{1}{\sqrt{4\pi}}$.

order invariants given as follows:

$$\hat{W}_{l,i} = \frac{\sum_{m_1}\sum_{m_2}\sum_{m_3}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix}\bar{Q}_{l,m_1,i},\bar{Q}_{l,m_2,i},\bar{Q}_{l,m_3,i}}{(\sum_m |\bar{Q}_{l,m,i}|)^{3/2}} \tag{6}$$

where $\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix}$ denote the Wigner 3-J symbols and they arise as the maximally symmetric coefficients of three coupled spherical harmonics with vanishing total angular momentum [51]. This way the $m$ dependency is avoided and the resulting expression is rendered invariant under arbitrary reorientations of the coordinate system. Third-order invariants, in their non-normalized form, were first explicitly constructed in the Landau expansions of liquid to nematic transition [56]. Later, their normalized form is shown to be central for resolving major symmetries [55]. Jarie gives a complete account of the third-order invariants in relation to long-range order phenomena within the context of the Landau theory [57].

One remarkable property of these invariants is the persistent sense of parent lattice when computed for crystalline systems. For the three cubic systems; simple cubic (SC), body-centred cubic (BCC) and face-centred cubic (FCC), all invariants of all $l$ coincide up to a sign, whereas hexagonal close packed and icosahedral orders are manifested with different shape spectra in $\hat{W}_{l,i}$ [55]. Therefore, these parameters are expedient tools to explore the discerning features of the energy landscape. Most notably, the choice of $l = 6$ yields an icosahedral BOO parameter for flat space.

To establish one to one correspondence between two structures, invariants of all orders and of all $l$ values have to match [55]. The basic difference between different $l$ values is the allowable symmetries that can be resolved [55, 57]. In practice, comparison of the first few degrees is sufficient to see if there is meaningful similarity between two structures. In the current work, we use the second-order invariants of $Q_{l,i}$ and the third-order invariants $\hat{W}_{l,i}$ with $l = 2$, 4 and 6.

Steinhardt *et al.* offered BOO as a means for 'shape spectroscopy', and conventionally, BOO parameters constructed with $l = 6$ were used in the literature because it is the lowest $l$ that can distinguish cubic and hexagonal configurations as well as the non-crystallographic icosahedral symmetry. Whether a given $l$ representation can distinguish a given symmetry follows from the isotropy groups of O(3) [57]. Isotropy group comprises the elements of O(3) that leaves the $(2l+1)$-dimensional vector $\left[Q_{l,-l,i}, Q_{l,-l+1,i}, \ldots, Q_{l,l,i}\right]$ invariant. If no such group exists for a given $l$ the corresponding order parameters $Q_{l,i}$ and $\hat{W}_{l,i}$ will be identically zero.

2.2.2 *Network analysis* Network view provides a convenient framework for comparing two systems in terms of inter-unit interactions. In the current study, networks are constructed from the residues of the generated structures by connecting residues that were within a 7 Å cut-off distance of each other. The choice of cut-off distance is motivated in part because 7 Å allows non-bonded interactions within the two nearest coordination shells to be included. This cut-off value is also consistent with previous Delaunay tessellation and cut-off scanning studies [58, 59].

The contact topology of a residue network is stored in the $N \times N$ adjacency matrix $A$, where the $A_{i,j}$ term is either equal to one or zero according to the presence or absence of interactions between the $i$th and $j$th residues, respectively [60]. Analysis of the adjacency matrix and its variants (such as the Laplacian or the normalized Laplacian) leads to a wealth of information. For example, a statistical mechanical treatment yields information on residue auto- and cross-correlations [61], and the neighbourhood structure of a

residue and how information propagates to further neighbours [62]. Network models were also used to identify adaptive mechanisms in response to perturbations [63], predict collective domain motions, hot spots and conserved sites [61, 64–66].

In terms of network characteristics, proteins share common structural similarities with other self-organizing condensed matter systems [62]. A quantity termed 'relative contact order', which may be derived from the adjacency matrix, is shown to highly correlate with the folding rates measured for many two-state folders [67].

In the current study, distributions and averages of local network parameters such as degree ($k_i$, number of neighbours of residue $i$, Eq. (7)), the nearest neighbour degree ($k_{nn,i}$, the number of neighbours of the neighbours of residue $i$, Eq. (8)) and the clustering coefficient ($C_i$) of residue $i$, Eq. (4), are monitored. In addition, a global parameter, the shortest path length ($L_i$) of residue $i$ is also monitored (Eq. (10)). These network parameters are defined as follows:

$$k_i = \sum_{j=1}^{N_i} A_{i,j} \tag{7}$$

$$k_{nn,i} = \frac{\sum_{j=1}^{N_i} \sum_{l=1}^{N_i} A_{i,j} A_{j,l}}{k_i} \tag{8}$$

$$C_i = \frac{\sum_{j=1}^{N_i} \sum_{l=1}^{N_i} A_{i,j} A_{j,l} A_{l,i}}{\frac{k_i(k_i-1)}{2}} \tag{9}$$

$$L_i = \frac{1}{N_i - 1} \sum_{i \neq j} L_{i,j} \tag{10}$$

For proteins and other self-assembled systems, the nearest neighbour degree ($k_{nn}$) is an important classifier and is a linearly increasing function of degree ($k$) with a slope that is equal to the average clustering coefficient ($\bar{C} = N^{-1} \sum_{i=1}^{N} C_i$) [62]. The clustering coefficient ($C_i$) is indicative of the amount of local structure present in the immediate neighbourhood of a residue. The clustering coefficient has been shown to converge to a value of approximately one-third for residues that are buried in the protein [60]. In Eq. (10), $L_{i,j}$ is the minimum number of steps between residues $i$ and $j$. The average shortest path length $\bar{L}$ calculated from $L_i$ values is a global parameter displaying the efficiency with which different parts of a protein communicate with each other and is shown to highly correlate with residue fluctuations [60].

The Laplacian ($L$) is used extensively in the graph theory literature [68, 69] and is defined as follows:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}, \tag{11}$$

where $\boldsymbol{D}$ is a diagonal matrix with $D_{i,i} = k_i$. The Laplacian is a positive-semi-definite matrix and its eigenvalue spectrum is routinely used for quantitative characterization of networks. In the current study, the normalized Laplacian ($L^*$) is used.

$$\boldsymbol{L}^* = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2}. \tag{12}$$

The main utility of the normalized Laplacian comes from the fact that all of its eigenvalues exist within the (0–2) interval [70, 71]. Therefore, the spectra of networks of differing sizes can be directly compared if the Laplacian matrix is normalized. More specifically, the presence of an eigenvalue with $\lambda = 2$ signals

that the graph has a bipartite connected component. The presence of an eigenvalue with $\lambda = 1$ shows that there are two vertices with identical connections (vertex doubling); and more generally, the multiplicity of eigenvalues with $\lambda = 1$ quantifies the extent of replication among structural motifs. The number of eigenvalues with $\lambda = 0$ is equal to the number of connected components. Finally, the smallest non-zero eigenvalue can be used as a quantitative measure of the difficulty to disconnect the network (the greater the value of the first non-zero eigenvalue is, the harder to disconnect the network becomes), and therefore, is a measure of collectivity in the network. Here, a connected component refers to a set of nodes for which there is at least one path among any pair of nodes in the set [69].

Spectral properties of the adjacency matrix can also be used to classify complex networks into four universal classes in terms of their topological structure. This classification is based on the spectral deviations from ideal networks for which a linear analytical relationship exists between the elements of the Perron–Frobenius eigenvector $\gamma_l^{\text{ideal}}$ of the adjacency matrix $A$, its eigenvalues $\lambda_i$ and the odd subgraph centrality ${}^{S}C_{\text{odd}}\,(i)$ of the node $i$ in the constructed network [72]:

$$\log \gamma_l^{\text{ideal}}\,(i) = 0.5 \log {}^{S}C_{\text{odd}}\,(i) - 0.5 \log[\sinh(\lambda_i)]. \tag{13}$$

For a network with $N_+$ nodes showing positive and $N_-$ nodes showing negative deviations from this perfect line, a total measure can be defined for each case by adding contributions from individual nodes as follows [72]:

$$\xi^+ = \sqrt{\frac{1}{N_+}\sum_{+}\left(\log\frac{\gamma_1(i)}{\gamma_1^{\text{ideal}}(i)}\right)^2} \quad \text{and} \quad \xi^- = \sqrt{\frac{1}{N_-}\sum_{-}\left(\log\frac{\gamma_1(i)}{\gamma_1^{\text{ideal}}(i)}\right)^2}. \tag{14}$$

Different topological structural classes can be clearly identified on a plot of $\xi^+$ against $\xi^-$. In the four-class designation Class I represents the case with perfect spectral scaling which is brought about by homogeneously distributed links, Class II represents the cases with negative deviations, Class III represents the case with positive deviations and finally Class IV represents the case with mixed deviations.

2.2.3 *Anistropic network model* ANM is an elastic network model, and it offers a physically grounded route to compute vibrational properties on coarse-grained systems with minimal input [3]. With the ANM method, for a system of $N$ particles, the Hessian matrix ($\mathbf{H}$) can be directly constructed from the spatial distribution of neighbour contacts by invoking a detailed force balance around each constituent particle. For a system which comprises $N$ particles and $M$ interactions, $\mathbf{H}$ is then given as:

$$\mathbf{H}_{3N\times 3N} = \gamma \mathbf{B}_{3\text{N}\times\text{M}} \mathbf{B}^T_{M\times 3N}. \tag{15}$$

Here, $\mathbf{B}_{3\text{N}\times\text{M}}$ stores the direction cosine vectors of the bonds connecting the particles and $\gamma$ is the effective force constant. Hessian matrix naturally encodes the contact topology of the whole network as well. Once the Hessian matrix is constructed from the particle coordinate data, the vibrational modes and the associated frequencies can be studied through its eigenvectors and the eigenvalue spectrum, respectively. In the present study, we work with the eigenvalues $\lambda_{i,\text{ANM}}$ of the $\gamma^{-1}\mathbf{H}_{3N\times 3N}$, which can be converted to the corresponding vibrational frequencies if the effective force constants are accounted for. Effective force constants can be extracted for each protein from the experimentally measured temperature factors but with artificial structures this is not possible. Since the conversion between $\lambda_i$ and the vibrational frequencies is only a rescaling operation ($\omega_i = \sqrt{\gamma\lambda_{i,\text{ANM}}}$), it is sufficient for us to use the $\lambda_i$ spectra for each protein to compare the ensemble distributions.

## 3. Results and discussion

Throughout this study, we will show that at least three simple requirements are necessary to generate an artificial structure (from a crystalline template) that mimics the behaviour of average folded protein:

(1) The presence of a molecular surface (finite size effect)

(2) The inclusion of free space

(3) Positional disorder

Furthermore, it will be shown that although the third requirement is important in reproducing the average protein radial distribution function, it has almost no impact on a variety of other structural features observed in real folded proteins.

As mentioned earlier, instead of targeting a specific protein to mimic (or identifying) its structure, we target the average behaviour of many proteins based on the experimentally determined coordinates. Therefore, as a first step, it is necessary to define a set of proteins ('the basis set') from which the average structural features of real proteins can be defined and the generated structures can be compared against.

### 3.1 *The average protein structure and the selection of lattice template*

To define an average protein structure, a set of 210 globular proteins was selected from 595 proteins, which are representative of four major protein folds with less than 25% sequence homology among them [60, 73]. The selection of the basis proteins was decided according to their globularity, which is investigated by a principle component analysis of amino acid $C_\alpha$ coordinates of these 595 proteins via singular value decomposition [74]. The proteins, whose ratio of the largest and lowest singular values ($\sigma_{\max}/\sigma_{\min}$) is less than 1.8 were selected into the basis protein set. The ratio of 1.8 is based on hen egg-white lysozyme (PDB Code: 193L), which is one of the most studied proteins. This selection criterion provided 210 proteins after eliminating rod-like proteins from consideration. The details of the basis set proteins are provided in the Supplementary Data section.

We first compare defect-free crystalline systems using BOO to see whether the local coordination characteristics in proteins can be directly reproduced by any of them. The average BOO parameters, $\bar{Q}_{l,i}$ and $\widehat{W}_{l,i}$, that were calculated from the 210 basis protein structures are shown in Fig. 1.

The average $\bar{Q}_{l,i}$ and $\hat{W}_{l,i}$ values suggests that the average residue neighbourhood is not a direct analogue of crystals [55]. Icosahedral order is also deemed to be absent for similar reasons, because icosahedral order requires that $\bar{Q}_{l,i} = 0$ for $l = 2, 4$, and this is far from being the case as can be seen in Fig. 1. Since comparisons with ideal crystalline lattices fail, one should consider drawing comparisons to systems that contain defects such as point defects (voids) or positional disorder.

RDF is an important metric of any packing consideration; therefore, any attempt to involve packing arguments should naturally start by an investigation of the RDF. The RDF of proteins (coarse-grained at the amino acid level) is known to have generic characteristics [75, 76]. The individual RDFs (grey-shaded region) and the average RDF (solid line) of the 210 basis proteins are shown in Fig. 2.

The average RDF is marked by a sharp first coordination peak at 3.8 Å and is the due to the bonds between nearest neighbours on the main chain. In addition, a less prominent secondary peak is observed at ∼5.6 Å. The ratio of the first and second peak locations puts fundamental constraints on the placement of neighbouring residues on a lattice. Fortunately, for both the FCC and SC lattices, the ratio of the first and second nearest neighbour distances is equal to $\sqrt{2}$ (1.4142), which is quite close to the ratio obtained from the average protein RDF (5.6/3.8 = 1.4737). Therefore, it is plausible to accept FCC and SC lattices as

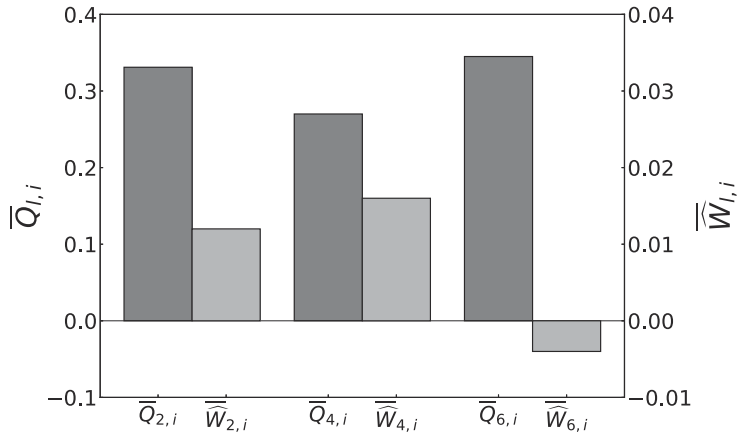


FIG. 1. Average BOO parameters, $\bar{Q}_{l,i}$ (dark grey) and $\widehat{W}_{l,i}$ (light grey), for $l = 2$, 4 and 6 for the 210 globular proteins used as the basis set.

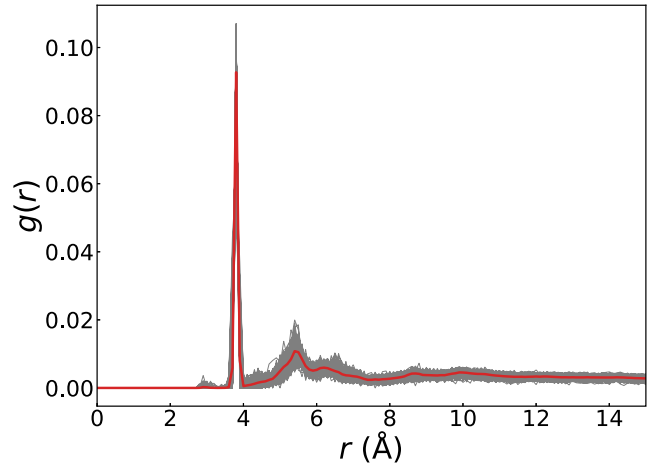


FIG. 2. Radial distribution functions, $g(r)$, of 210 globular proteins (thin grey lines) and their average radial distribution function (bold line) calculated with a bin size of 0.1 Å.

candidate template lattices. However, when the average number of nearest and next-to-nearest neighbours of nodes in proteins is calculated from the area underneath each peak, it becomes evident that the first coordination shell needs to be less crowded than the second coordination shell. However, in FCC, which is a close-packed structure, the opposite is true and therefore, the FCC lattice cannot be considered as a template lattice. This leaves the SC lattice as the only candidate. Although this selection might be viewed as counterintuitive, it has been shown that the bending and torsional angles of proteins coarse-grained at the amino acid level show peaks at highly preferential values agreeing with the geometry of the SC lattice [36]. Therefore, the selection of SC lattice as the template for the generation of coarse-grained, artificial protein structures is not a coincidence, and it is based on multiple structural considerations for real proteins.

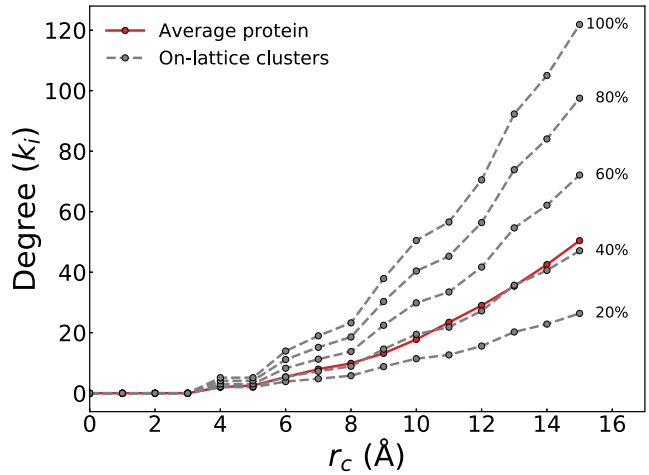


FIG. 3. The dependence of the average degree ($k$) on cut-off distance for 210 basis proteins (solid line) and SC clusters (dashed lines) at various occupancies.

### 3.2 *Generation of finite-sized clusters*

The first step in the generation of artificial proteins is the carving of finite-sized clusters from the SC lattice template. This step is required to capture the finite size of proteins. However, it is clear that any fully occupied finite-sized SC cluster would be too dense compared to proteins. The cut-off variation of the average contact numbers (degree, $k$) for the 210 basis proteins and various SC clusters as a function of occupancy are shown in Fig. 3, which clearly shows that the perfect SC cluster at 100% occupancy is indeed too dense compared to the average behaviour of real proteins.

Interestingly, at 40% occupancy (after randomly removing 60% of the lattice points), the average contact density of 210 SC clusters shows almost the same cut-off variation as that of the average protein. This finding suggests that in order to create structures similar to proteins, the SC clusters should be emptied down to ∼40% occupancy. We also note that 40% occupancy as well as the 60% void concentration are both above the percolation threshold of 25% of the SC lattice [77]. This suggests that once the SC cluster is emptied down to 40% occupancy, the remaining residues could form a connected network of points, which is an absolute requirement for any protein-like structure to be considered viable. However, random removal of lattice sites does not guarantee that the remaining lattice sites form a connected network; therefore, an additional constraint is imposed during the random removal procedure to maintain the connectedness of the remaining lattice points. In the current study, the connectedness is defined such that there must be at least one path between any two remaining beads on the cluster.

Finally, three-dimensional networks are constructed by introducing additional bonds among the constituent beads that account for the second coordination shell. The construction process is illustrated on a $7 \times 7 \times 7$ parent cluster in Fig. 4. Since the residues of the structures generated on the SC lattice template are located on regularly spaced lattice positions, their RDF would show up as a succession of delta functions with no spread irrespective of occupancy, see Fig. 5(a).

It is well known that positional disorder (displacement of lattice points from their ideal locations) creates spread in RDF peaks. Therefore, the random displacement and the RMC–SA procedures are used to evolve the (emptied) clusters.

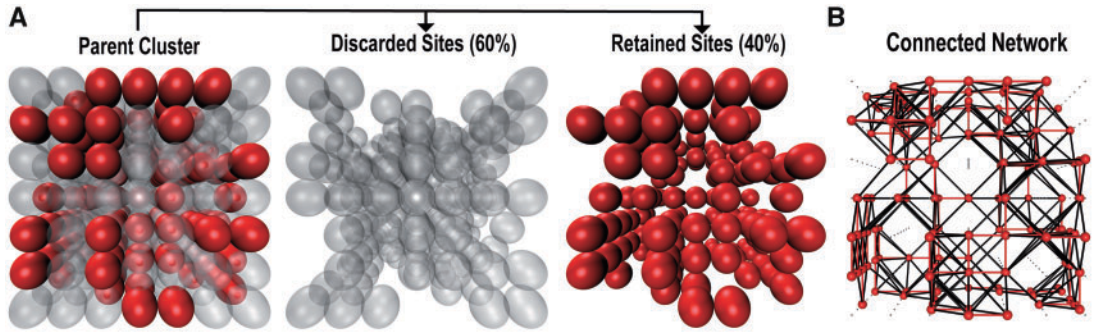


FIG. 4. Construction of a three-dimensional artificial network on a parent $7 \times 7 \times 7$ SC cluster. (a) Parent cluster is emptied down to 40% occupancy by discarding 60% of the initial sites. The remaining structure comprises 138 beads which form a single connected component. This portion was rendered with VMD [78]. (b) An example network construction is shown by adopting a 7 Å cut-off radius. In this figure, the nearest neighbour contacts are drawn in red whereas additional contacts appear in black. This portion was rendered with Vesta [79]. For the analyses, no distinction is made between the nearest neighbour contacts and the non-nearest neighbour contacts similar to the treatment of nodes in ANM as applied on real proteins.

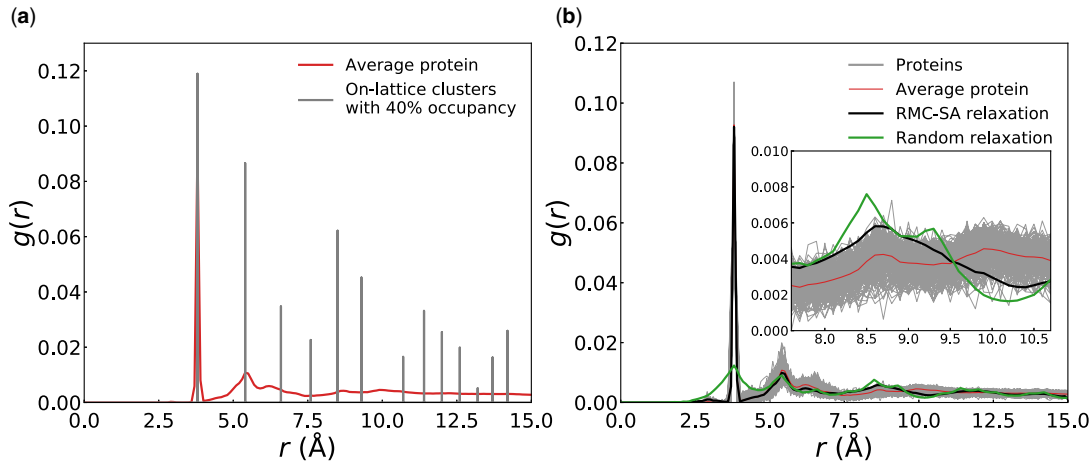


FIG. 5. (a) Radial distribution functions, $g(r)$, of the average protein and an example on-lattice cluster structure at 40% occupancy before introducing positional disorder. (b) Comparison of the average radial distribution functions of 210 basis proteins and 210 clusters at 40% occupancy after relaxation of the on-lattice residue locations with random displacements (dashed line) or RMC–SA procedure (solid line) calculated using a bin size of 0.1 Å. Inset shows the region of the largest discrepancy between real proteins and generated structures.

### 3.3 *Introduction of positional disorder*

The comparison of the average RDF for 210 basis proteins and 210 clusters at 40% occupancy relaxed with random displacements or with RMC–SA procedure is shown in Fig. 5(b). The relaxation procedure performed with random displacement of residues does not reproduce the average RDF of the basis proteins, the main difference between the two RDFs being at the first coordination peak. On the other hand, because the RMC–SA relaxation procedure uses the average protein RDF as its target function, the RDF of the cluster relaxed with RMC–SA procedure successfully matches that of the average protein.

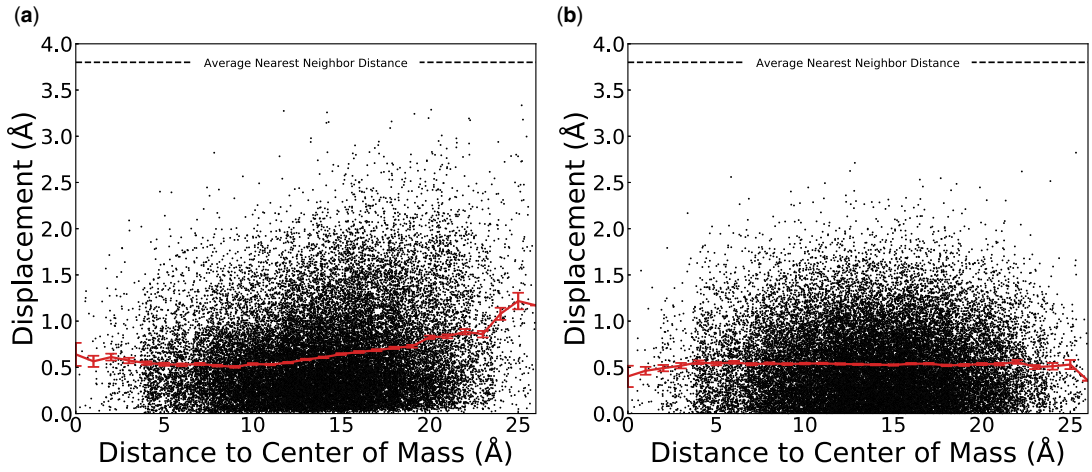


FIG. 6. Distribution of the radial displacements of all simulated residues as a function of radial distance from the centre of mass of the cluster during (a) RMC–SA and (b) random displacement procedure. Solid curves are moving averages and the error bars represent the standard error.

The main discrepancy between the two RDFs is around the distance range of 7.6–10.7 Å (see inset to Fig. 5(b)). The sub-range from 7.6 Å to 9.3 Å corresponds to three on-lattice peaks that are representative of the 4th, 5th and 6th closest neighbours on the SC lattice (Fig. 5(a)), and it is clear that the RMC–SA relaxation procedure is not able to reduce the high number of initial contacts that were present in the on-lattice structure within this distance range. The reasons for this outcome are not known. However, one possible explanation might be that the connectedness constraint imposed during the introduction of voids is not strict enough to produce a linear chain, and therefore, side branches materialize after the introduction of voids. Because the relaxation procedure does not remove branches, their existence leads to a more compact structure and a slightly higher density within the core of the cluster. From 9.3 Å to 10.7 Å, the source of discrepancy is more straightforward. This range is the empty region between the 6th and 7th nearest neighbours on SC lattice (see Fig. 5(a)), and it is larger than the average displacements introduced with the two relaxation methods. Therefore, RDF functions pertaining to the generated structures attain lower values in comparison to the proteins.

The radial displacements of all residues and the average radial displacement as a function of distance from the centre of mass are shown in Fig. 6 for the two relaxation procedures employed. Neither the individual nor the average radial displacements are ever greater than the average nearest neighbour distance (3.8 Å) between residues regardless of the location of the residue from the centre of mass (buried at the protein core or at the free surface). However, the residues at the free surface experience (on average) slightly greater displacements compared to those closer to the centre of mass during the RMC–SA relaxation, see Fig. 6(a).

When the relaxation is performed with completely random displacements, the difference between the core and surface disappears, see Fig. 6(b). Although the RMC–SA relaxation procedure does not inherently impose any location-specific constraints, it is clear that the unintended consequence of trying to fit the average RDF of the basis proteins (which is the target function of the RMC–SA procedure) leads to this outcome. This behaviour is also compatible with the extra mobility that would be enjoyed by residues residing on the surface of real proteins.

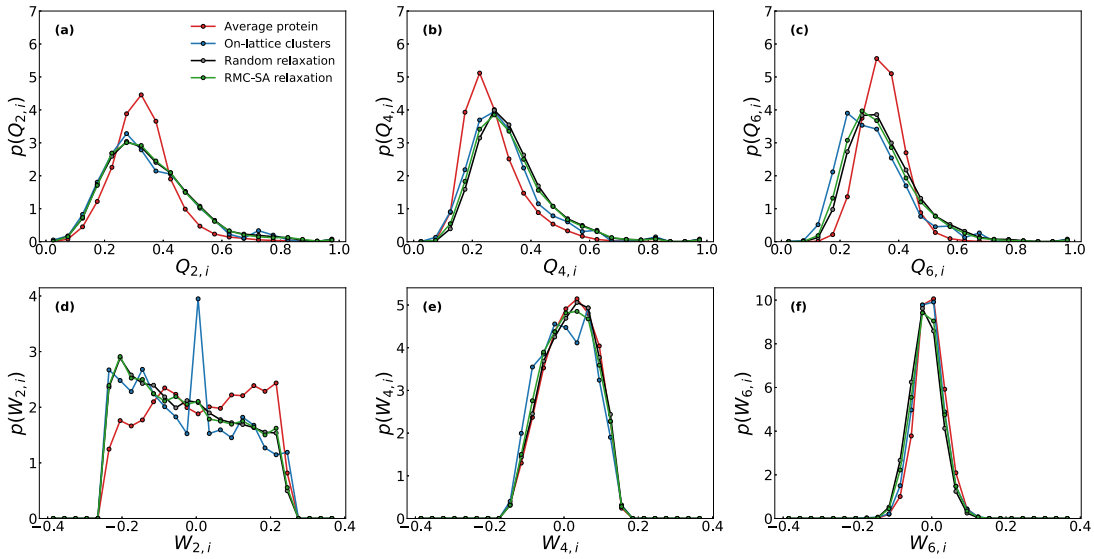


FIG. 7. Comparison of various BOO parameter distributions of average protein and various generated structures with 40% occupancy for $l = 2$, 4 and 6.

## 3.4 *Comparison of structural features*

3.4.1 *Bond orientational order analysis* RDF is a reduced (directionally averaged) pair distribution function. Although it can distinguish different condensed phases of matter such as crystals, liquids and gases, a multitude of structures could realize the same RDF. Therefore, an optimization which only takes the RDF as its objective function is not guaranteed to generate a unique structure. To understand if the generated structures are comparable to real proteins, one needs to match other properties. For this reason, additional comparisons are performed both geometrically and topologically. The geometric comparisons are based on BOO and the relevant order metrics $\bar{Q}_{l,i}$ and $\widehat{W}_{l,i}$ for $l = 2, 4$ and 6. BOO parameters quantify the differences in anisotropy of the local environment. All BOO parameters are computed with a cut-off distance of 7 Å. This choice is based on previous Voronoi studies which showed that a nearest neighbour definition should not be limited to the intra-chain neighbours (that are separated by 3.8 Å) only and that all neighbours within 7 Å should be considered [24, 56]. In addition, this choice of cut-off distance accounts for non-bonded interactions within the first two neighbour shells. The results of the BOO analysis are shown in Fig. 7 and Table 1. There is nontrivial correspondence between BOO parameters computed for all degrees of $l$. For each $l$ studied, on-lattice and relaxed clusters produce distributions that are similar to that of the average protein in terms of shape (unimodal distributions) and range/variation. In general, the real proteins produce $Q_{l,i}$ distributions with stronger peaks. For $l = 2$ and 4, the average protein has slightly lower mean values for $Q_{l,i}$. This is most likely a manifestation of an underlying chain, which puts restrictions on the organization of non-bonded contacts. Additionally, for $Q_{2,i}$ parameter, the fact that it cannot resolve cubic symmetry might contribute to the observed discrepancy.

Finally, the BOO parameters have a size component which scales inversely with the contact number, and because proteins have a slightly higher average contact number, the protein distributions might translate to slightly lower average BOO parameters although this is not observed for $Q_{l,i}$ for $l = 6$.

TABLE 1 *Mean values of various Bond Orientational Order and network parameters for 210 basis proteins, unrelaxed on-lattice clusters, clusters relaxed with random move algorithm and clusters relaxed with RMC–SA algorithm. All generated structures have 40% occupancy. The average eigenvalue of the normalized Laplacian is not shown because it is always equal to one by definition. Standard error of the last digit is shown in parentheses only when it is greater than 1%*

| Parameter | Protein | Unrelaxed (on-lattice) | Relaxed (random) | Relaxed (RMC–SA) |
|---|---|---|---|---|
| $\bar{Q}_{2,i}$ | 0.331 | 0.343 | 0.352 | 0.350 |
| $\bar{Q}_{4,i}$ | 0.270 | 0.310 | 0.334 | 0.327 |
| $\bar{Q}_{6,i}$ | 0.345 | 0.317 | 0.349 | 0.338 |
| $\overline{\hat{W}}_{2,i}$ | 0.012 | −0.028 | −0.030 | −0.030 |
| $\overline{\hat{W}}_{4,i}$ | 0.016 | 0.005 | 0.015 | 0.012 |
| $\overline{\hat{W}}_{6,i}$ | −0.004 | −0.009 | −0.017 | −0.013 |
| $\bar{k}$ | 7.947 | 7.381(3.6) | 7.162(3.4) | 7.117(3.1) |
| $\bar{k}_{nn}$ | 8.344 | 8.080(3.5) | 7.799(3.4) | 7.739(3.1) |
| $\bar{C}$ | 0.550 | 0.547 | 0.542 | 0.535 |
| $\bar{L}$ | 5.129(6.7) | 4.199(4.8) | 4.225(4.7) | 4.220(4.6) |
| $\lambda_{\mathrm{ANM}}\,(7\text{Å})$ | 2.649 | 2.460 | 2.387 | 2.372 |
| $\lambda_{\mathrm{ANM}}\,(10\text{Å})$ | 5.924 | 6.575 | 6.459 | 6.292 |
| $\lambda_{\mathrm{ANM}}\,(13\text{Å})$ | 11.794 | 12.077 | 11.988 | 11.774 |

In general the correspondence between the average protein and the relaxed structures is better for $\hat{W}_{l,i}$ than it is for $Q_{l,i}$ for $l = 4$ and 6. The (unrelaxed) on-lattice clusters show a strong peak at $\hat{W}_{2,i}$ (Fig. 7(d)) because $\hat{W}_{2,i}$ vanishes for $l = 2$ in a cubic cluster [55] and the starting structure will have points that sit at an ideal cubic environment. Although this peak quickly disappears once the structure is relaxed through random moves and RMC–SA algorithm, the relaxed probability distribution functions show the opposite variation in $\hat{W}_{2,i}$ than that is observed for the average protein structure. This discrepancy seems to be the most prominent disagreement between the proteins and relaxed structures.

It is clear that although RMC–SA algorithm helps us to recover the spread in the RDF function, it has much less effect on the BOO parameters due to the small magnitude of displacements introduced. Relaxation with random moves gives almost the same BOO parameter distributions as those obtained with RMC–SA algorithm. In fact, with the exception of the sharp peak observed at $\hat{W}_{2,i} = 0$, the on-lattice structures show distributions quite similar to those of the relaxed structures. This is an important finding in the sense that the template lattice (at 40% occupancy) seems to provide most of the necessary structural similarities. Although structurally neither the on-lattice nor the relaxed structures are exactly the same as the average protein structure, the fact that they reproduce most of the main features of the measures used is remarkable. We also note that the generated structures could be improved if more rigorous optimization methods are employed. For example, the RMC–SA algorithm could be made to optimize against a collection of structural properties in addition to the RDF, and as a result, better agreement could be obtained between the structural features compared. However, because the current study aims to identify the simplest set of rules to generate structures similar to folded proteins, a more rigorous approach is not pursued at this time.

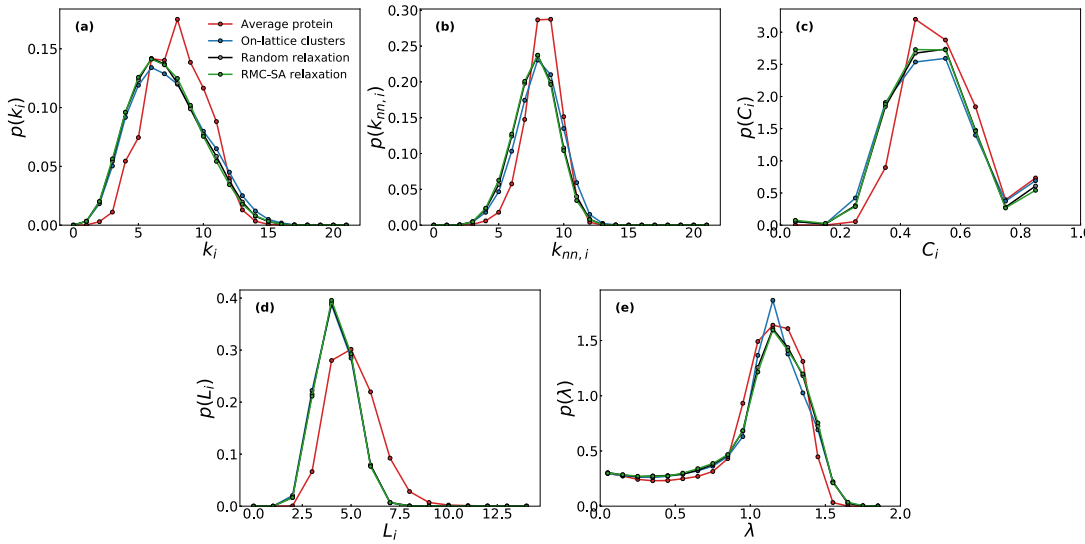


FIG. 8. Comparison of the distributions of the (a) degree (contact number), (b) nearest neighbour degree, (c) clustering coefficient, (d) shortest path length and (e) eigenvalues of the normalized Laplacian for average protein and generated clusters with 40% occupancy.

3.4.2 *Network analysis* The probability distributions of various network parameters of the average protein and various generated structures are presented in Fig. 8 (mean values of these parameters are tabulated in Table 1). Comparison of the degree ($k_i$) distributions suggests that the proteins have slightly greater contact number within the 7 Å cut-off distance, see Fig. 8(a). This observation may be attributed to the fact that the coarse-grained protein is a self-avoiding chain, whereas the connectivity criteria used during the generation of the clusters do not impose a self-avoiding walk constraint. The effects of the connectivity criteria are also seen in the nearest neighbour degree ($k_{nn}$), clustering coefficient ($C$) and shortest path length ($L$) distributions (Fig. 8(b–d), respectively). However, despite the slight differences in the local and global network parameters, the generated clusters capture the network-specific properties of proteins considerably well.

Figure 8(e) displays the eigenvalue distribution of the normalized Laplacian for the generated clusters and the average protein. The eigenvalue distribution of the average protein is characterized by a wide peak in the region $1.0 < \lambda < 1.5$, overlaid by a long tail at $0 < \lambda < 1.0$. The eigenvalues in the proximity of $\lambda = 0$ carry signatures of collectivity in the system and indicates the ease with which the network may be disconnected; the generated structures capture this region remarkably well. In addition, the region where $\lambda > 1$ is marked by a wealth of local motifs that are realized by secondary structural elements, and therefore, it is expected to cause the biggest discrepancies. However, the generated clusters successfully mimic the behaviour of the average protein structure in the region $\lambda > 1$. Finally, the overall distribution of the generated clusters is observed to be typical of the molecular structure networks [62, 80].

One final test of topological similarities between globular proteins and the on-lattice structures is based on the spectral scaling properties (Eqs 13 and 14). Based on positive and negative deviations from ideal spectral scaling line, there arise four universal topological classes. Estrada shows that proteins typically populate the Class II [47]. This class is characterized by high modularity which is attained by densely connected clusters demarcated with structural cavities. This work also shows that major network

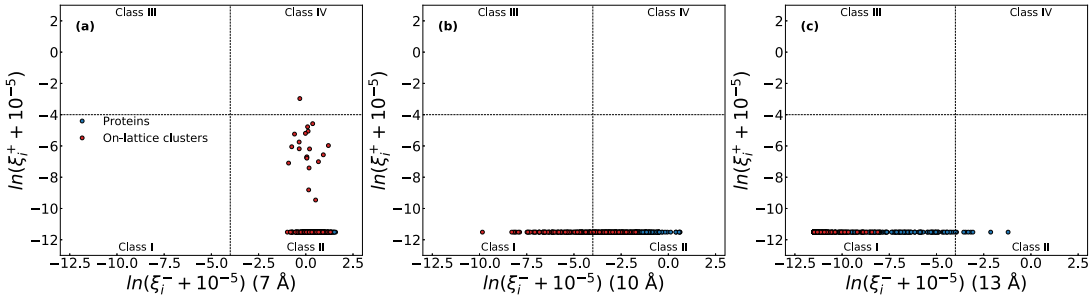


FIG. 9. Comparisons of proteins with artificial structures in terms of spectral deviations from ideal networks at three different cut-offs 7 Å, 10 Å and 13 Å. Vertical and horizontal dashed lines dissect the graph into four universal topological structural classes. Structures for which all the constituent nodes show negative deviations are automatically assigned $\xi_i^+ = 0$. On the graph these fall onto the same horizontal line with the ordinate $\ln\left(10^{-5}\right)$ since all the values are augmented with $10^{-5}$ before natural log transformation.

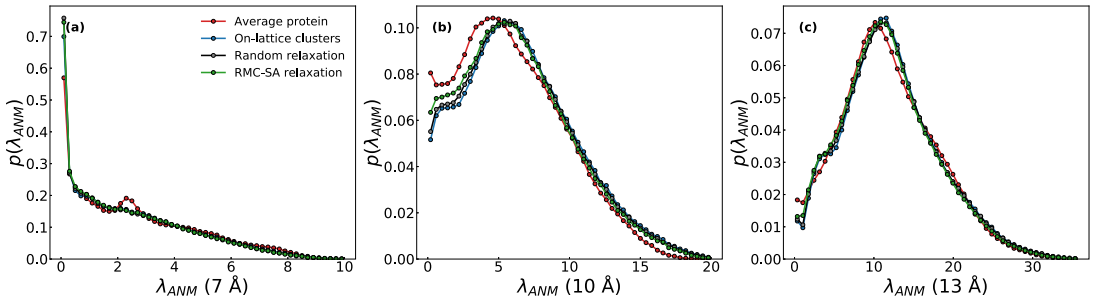


FIG. 10. $\lambda^{-1}\mathbf{H}$ spectra comparisons based on the ANM method are shown at three different cut-offs. At lower cut-offs there are more than six zero eigenvalues because force balance condition cannot be satisfied around each node. With increasing cut-off a more symmetric Gaussian-like distribution is recovered.

growth models such as Erdös–Rényi and Barabási–Albert cannot produce structures in this topological class [47]. In Fig. 9, we show the comparisons between globular proteins and on-lattice structures at three different cut-off values on a plot of $\xi^+$ against $\xi^-$ defined in Eq. 14.

It is noteworthy that our network generation model can automatically reproduce the topological characteristics of Class II networks with a minimalistic set of rules. Moreover, the transition from Class II to Class I with increasing cut-off value is captured, as well. Class I describes networks with ideal spectral scaling and free of communication bottlenecks due to highly regular and homogenized link distributions. This cut-off dependent transitive behaviour towards ideal scaling is more pronounced for on-lattice structures; mostly likely as a manifestation of the underlying cubic crystalline order.

3.4.3 *Vibrational characteristics through ANM* Vibrational characteristics of globular proteins have well-known features as manifested in the distribution of modal frequencies which possesses Gaussian characteristics [81, 82]. To extend our comparisons beyond static structural descriptors, we finally compare the vibrational spectra of the two ensembles through ANM in Fig. 10.

Our results show that the artificially generated models successfully capture the spectral distributions observed from the networks constructed from real proteins. By holding comparisons at 7, 10 and

13 Å cut-offs, we show that the observed correspondence is independent of the parameterization of the ANM. With increasing cut-off distance, the Gaussian characteristic of globular proteins is recovered, as well. This finding hints that the structural similarities are accompanied by similarities in conformational flexibility under a generic harmonic potential.

3.4.4 *Statistical comparison of distributions* To quantify the extent of similarity between the generated structures and the average protein structure, two-sample Kolmogorov–Smirnov tests are performed. The Kolmogorov–Smirnov test is a non-parametric statistical test that compares two independent distributions and is sensitive to location and shape of distributions. The Kolmogorov–Smirnov test calculates the maximum distance ($D$) between two cumulative distribution functions (CDFs). In our case, we compared the generated structures to real proteins; therefore, $D_x$ is calculated for various properties ($x$ stands for various BOO or network parameters investigated) between the average cumulative distribution functions of 210 proteins and 210 generated structures as follows [83]:

$$D_x = \max \left| \mathrm{CDF}_{\mathrm{proteins},x} - \mathrm{CDF}_{\mathrm{structures},x} \right| . \tag{16}$$

In the two-sample Kolmogorov–Smirnov test, the null hypothesis is rejected with a confidence level ($\alpha$) if the following inequality holds:

$$D_x > d_\alpha \sqrt{\frac{\mathrm{N}_{\mathrm{proteins}} + \mathrm{N}_{\mathrm{structures}}}{\mathrm{N}_{\mathrm{proteins}} \mathrm{N}_{\mathrm{structures}}}} \tag{17}$$

where $N_{\mathrm{proteins}}$ and $N_{\mathrm{structures}}$ are the number of data points making up the respective CDFs. The value of $d_\alpha$ (tabulated in [83]) depends on $\alpha$ (a value of 0.05 is used generally) and sample size. In the current study, the null hypothesis is that both groups were sampled from populations with identical distributions; therefore, we would like to see that this null hypothesis is not rejected. To provide a confidence level to (accepting or rejecting) the null hypothesis, we calculated $p$-values, which indicate the probability of obtaining a result equal to or more extreme than what is actually observed assuming that the null hypothesis is true. In general, the null hypothesis is rejected if the $p$-value is less than a predetermined confidence level (generally 0.05 or 5%).

The results of the two-sample Kolmogorov–Smirnov tests are shown in Table 2. It is clearly seen from the $p$-values that the null hypothesis (that the two samples, proteins vs. various generated structures, are drawn from identical distributions) is not rejected for any of the generated structures at 40% occupancy. As a result, it can be stated that the correspondence obtained between proteins and various generated structures is a statistically significant observation.

### 3.5 *Influence of occupancy*

The results of the structural comparison of the unrelaxed (on-lattice) and relaxed generated structures to the average of 210 real proteins reveal that with the exception of a few differences, the relaxation of the residues from their on-lattice locations do not result in substantial differences in the BOO and network properties. Thus, most of the structural features are inherited from the lattice template used to create these generated clusters. One of the rules imposed on the generation procedure is to empty the clusters such that the occupancy would be 40%. This leads to an excellent agreement between the average degree (contact number) of the on-lattice clusters and real proteins (see Fig. 3). However, it is important to understand the effect of occupancy on the structure of the generated clusters and their similarity to real protein structure.

TABLE 2 *Two-sample Kolmogorov–Smirnov test parameters comparing various structural property distribution functions of proteins with those of unrelaxed (on-lattice) and relaxed clusters. All generated structures have 40% occupancy*

| | Unrelaxed (on-lattice) | | Random relaxation | | RMC–SA relaxation | |
|---|---|---|---|---|---|---|
| Parameter | $D$ | $p$-value | $D$ | $p$-value | $D$ | $p$-value |
| $Q_{2,i}$ | 0.113 | 0.999 | 0.123 | 0.996 | 0.117 | 0.998 |
| $Q_{4,i}$ | 0.155 | 0.954 | 0.242 | 0.539 | 0.215 | 0.691 |
| $Q_{6,i}$ | 0.249 | 0.501 | 0.116 | 0.998 | 0.161 | 0.938 |
| $\hat{W}_{2,i}$ | 0.140 | 0.945 | 0.123 | 0.983 | 0.121 | 0.986 |
| $\hat{W}_{4,i}$ | 0.073 | 1.000 | 0.011 | 1.000 | 0.028 | 1.000 |
| $\hat{W}_{6,i}$ | 0.054 | 1.000 | 0.134 | 0.961 | 0.097 | 0.999 |
| $k_i$ | 0.139 | 0.974 | 0.154 | 0.937 | 0.158 | 0.924 |
| $k_{nn,i}$ | 0.117 | 0.997 | 0.176 | 0.848 | 0.189 | 0.783 |
| $C_i$ | 0.145 | 1.000 | 0.127 | 1.000 | 0.132 | 1.000 |
| $L_i$ | 0.283 | 0.291 | 0.276 | 0.319 | 0.277 | 0.315 |
| $\lambda$ | 0.047 | 1.000 | 0.049 | 1.000 | 0.053 | 1.000 |
| $\lambda_{\mathrm{ANM}}\,(7\text{Å})$ | 0.039 | 1.000 | 0.053 | 1.000 | 0.055 | 1.000 |
| $\lambda_{\mathrm{ANM}}\,(10\text{Å})$ | 0.068 | 1.000 | 0.056 | 1.000 | 0.040 | 1.000 |
| $\lambda_{\mathrm{ANM}}\,(13\text{Å})$ | 0.026 | 1.000 | 0.019 | 1.000 | 0.010 | 1.000 |

To investigate the effect of occupancy, we use the random relaxation procedure because there is very little difference between the random move algorithm and the more computationally demanding RMC–SA algorithm. Three occupancies are studied: 20, 40 and 60%. In the case of BOO parameters, increasing the occupancy leads to the narrowing of the $Q_{l,i}$ distributions and shifts them towards lower mean values (Fig. 11 and Table 3). For $l = 2$ and 6, $Q_{l,i}$ distributions of the generated structures at 40% occupancy match the protein distribution more closely than those at 20 and 60% occupancies. However, for $l = 4$, 60% occupancy distribution is the closest to that of the protein. These observations are also reflected in the Kolmogorov–Smirnov statistics (Table 4). The BOO parameter $\hat{W}_{l,i}$ is not sensitive to occupancy especially for $l = 4$ and 6. For example, the width of the $\hat{W}_{l,i}$ distributions do not change, in fact, the width does not change for any $l$.

On the other hand, for $l = 2$, the effect of occupancy is quite complex. The average protein structure shows an almost flat distribution around $\hat{W}_{2,i} = 0$ with a slight positive slope. On the other hand, the generated structures with 20% occupancy show a large negative slope. Increasing the occupancy increases the slope around $\hat{W}_{2,i} = 0$. At 60% occupancy, the $\hat{W}_{2,i}$ distribution of the generated structures become almost flat and is the closest to that of the average protein. This observation is supported by the Kolmogorov–Smirnov analysis, which also indicates that the 40% and 60% occupancies are statistically identical to each other and to the average protein (Table 4).

The degree and average neighbour degree show similar trends with respect to occupancy; the distributions of the generated structures become wider and shift towards greater mean values Fig. 12(a,b)). The clustering coefficient distributions show slight variations with respect to different percentage of occupancy; the distributions become slightly narrower with lower mean values upon increasing occupancy (Fig. 12(c)).

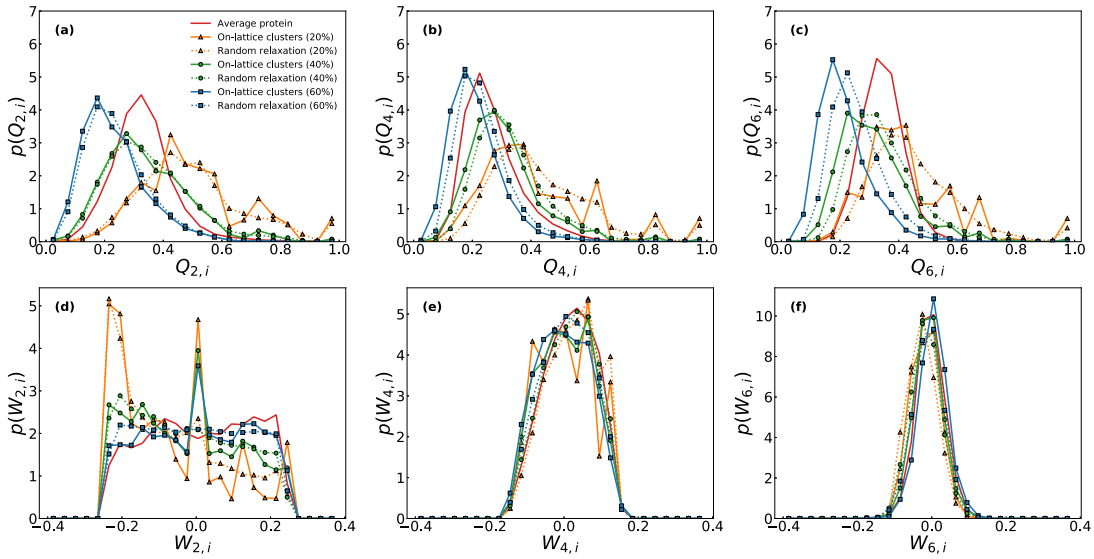


FIG. 11. Influence of occupancy on various BOO parameters for $l = 2$, 4 and 6.

TABLE 3 *Influence of occupancy on the mean values of various Bond Orientational Order and network parameters for unrelaxed on-lattice clusters and clusters relaxed with random move algorithm. The average eigenvalue of the normalized Laplacian is not shown because it is always equal to one by definition. Standard error is shown within parenthesis for the last digit and only when it is greater than 1%*

| | | Occupancy | | | | | |
|---|---|---|---|---|---|---|---|
| | | 20% | | 40% | | 60% | |
| Parameter | Protein | Unrelaxed (on-lattice) | Relaxed (random) | Unrelaxed (on-lattice) | Relaxed (random) | Unrelaxed (on-lattice) | Relaxed (random) |
| $\bar{Q}_{2,i}$ | 0.331 | 0.496 | 0.498 | 0.343 | 0.352 | 0.233 | 0.245 |
| $\bar{Q}_{4,i}$ | 0.270 | 0.419 | 0.441 | 0.310 | 0.334 | 0.213 | 0.237 |
| $\bar{Q}_{6,i}$ | 0.345 | 0.434 | 0.455 | 0.317 | 0.349 | 0.215 | 0.260 |
| $\overline{\hat{W}}_{2,i}$ | 0.012 | −0.071 | −0.071 | −0.028 | −0.030 | 0.005 | −0.004 |
| $\overline{\hat{W}}_{4,i}$ | 0.016 | 0.005 | 0.024 | 0.005 | 0.015 | −0.001 | 0.008 |
| $\overline{\hat{W}}_{6,i}$ | −0.004 | −0.018 | −0.025 | −0.009 | −0.017 | 0.003 | −0.009 |
| $\bar{k}$ | 7.947 | 4.811 | 4.717 | 7.381 | 7.162 | 11.275 | 10.919 |
| $\bar{k}_{nn}$ | 8.344 | 5.330 | 5.204 | 8.080 | 7.799 | 12.181 | 11.732 |
| $\bar{C}$ | 0.550 | 0.618 | 0.605 | 0.547 | 0.542 | 0.530 | 0.528 |
| $\bar{L}$ | 5.129 | 5.348 | 5.394 | 4.199 | 4.225 | 3.921 | 3.850 |

TABLE 4 *Influence of the occupancy on the two-sample Kolmogorov–Smirnov test parameters for structures relaxed with the random move algorithm. The on-lattice and RMC–SA relaxed results follow the same trends observed in the randomly relaxed structures. Italic text indicates p-values less than 0.05*

| | Occupancy | | | | | |
|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | |
| Parameter | $D$ | $p$-value | $D$ | $p$-value | $D$ | $p$-value |
| $Q_{2,i}$ | 0.488 | 0.011 | 0.123 | 0.996 | 0.389 | 0.071 |
| $Q_{4,i}$ | 0.481 | 0.012 | 0.242 | 0.539 | 0.152 | 0.961 |
| $Q_{6,i}$ | 0.381 | 0.083 | 0.116 | 0.998 | 0.460 | 0.019 |
| $\hat{W}_{2,i}$ | 0.243 | 0.376 | 0.123 | 0.983 | 0.049 | 1.000 |
| $\hat{W}_{4,i}$ | 0.052 | 1.000 | 0.011 | 1.000 | 0.046 | 1.000 |
| $\hat{W}_{6,i}$ | 0.221 | 0.493 | 0.134 | 0.961 | 0.073 | 1.000 |
| $k_i$ | 0.544 | 0.002 | 0.154 | 0.937 | 0.370 | 0.074 |
| $k_{nn,i}$ | 0.747 | 0.000 | 0.176 | 0.848 | 0.673 | 0.000 |
| $C_i$ | 0.118 | 1.000 | 0.127 | 1.000 | 0.088 | 1.000 |
| $L_i$ | 0.114 | 0.997 | 0.276 | 0.319 | 0.440 | 0.019 |
| $\lambda$ | 0.148 | 0.975 | 0.049 | 1.000 | 0.082 | 1.000 |

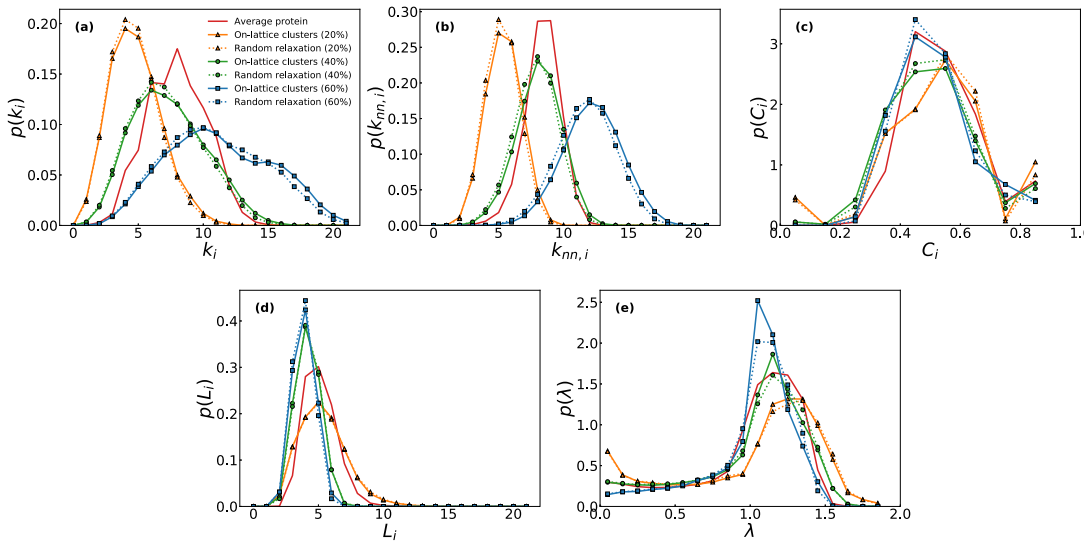


FIG. 12. Comparison of the distributions of (a) degree (contact number), (b) nearest neighbour degree, (c) clustering coefficient, (d) the shortest path length and (e) the eigenvalue of the normalized Laplacian for the average protein and generated clusters with respect to cluster occupancy.

The clustering coefficient distribution at 60% occupancy is closest to that of the average protein, although the Kolmogorov–Smirnov analysis did not reveal any statistical difference with respect to occupancy. This observation shows that a protein residue on average has a more densely connected

neighbourhood than the cubic cluster, a characteristic which is not fully governed by a global occupancy parameter that is used in the current study.

At 20 and 60% occupancy, the degree and nearest neighbour degree distributions of the generated structures show no statistical similarity to that of the average protein as suggested by the Kolmogorov–Smirnov analysis (Table 4). In both cases, 40% occupancy is the closest to the average protein. The relationship between the degree and occupancy is quite straightforward; the number of neighbours for any lattice site increases as a result of increasing occupancy. The same effect is also reflected in the nearest neighbour degree. It is obvious that a larger occupancy than 40%, but less than 60% is required to match the degree and average neighbour degree distributions of real proteins.

In the case of the shortest path length, the distributions of the generated structures shift towards lower mean values and become narrower with increasing occupancy (Fig. 12(d)). This is expected given that at greater occupancy, there is a greater possibility of finding a shorter path between any two residues. Examination of the Laplacian eigenvalue distributions (Fig. 12(e)) shows that the number of vertex doubling and the number of connected components increased with increasing occupancy. This is probably due to the connectivity constraint used during emptying of the lattice clusters which leads to a branched structure.

## 4. Conclusions

In this work, we contribute to the toolbox of methodologies aiming at characterizing as well as generating artificial structures that automatically reproduce the general structural characteristics of globular proteins. We show that simple cubic (SC) lattice is the proper structural template for such a construction. The relevance of cubic lattices has been originally motivated from the statistical distributions of internal coordinates defined by $C_\alpha$ atoms (i.e. $C_\alpha$–$C_\alpha$–$C_\alpha$ bending angle) [36]. In the present study, we show that defect laden SC arrangement is the proper cubic template and the closed packed FCC is incompatible with residue coordination characteristics. This observation is directly motivated from the average protein radial distribution function (RDF).

We identify three major sources of disorder that are sufficient to recover the geometric and topological properties of proteins with an artificial network structure if we start from a bulk SC crystal: the molecular (free) surface, random distribution of voids and small displacements from ideal lattice positions.

Molecular surface accounts for the finite size of protein molecules and is an obvious requirement. Cubic clusters are easy to generate and can conveniently replicate the compact shape characteristics of globular proteins. This choice is justified *a posteriori* as the shortest path lengths of generated structures match well with that of real proteins.

Our results single out randomly distributed voids as the primary source of disorder needed to reproduce the structural metrics of globular proteins. Most notably, the mean coordination number in the SC cluster mimics protein characteristics at an arbitrary cut-off only after introducing voids at a concentration of 60% into a perfect cluster. This observation is the core of our structure generation strategy which can automatically create structures that are in the same topological class as proteins and cannot be generated by network growth mechanisms such as Erdös–Renyi and Barabasi–Albert [72]. Estrada shows that protein residue networks belong to the Class II category which is characterized by modular structures where highly interconnected clusters are interspersed with structural voids [47]. It is noteworthy that with the cubic lattice background, the random introduction of voids becomes sufficient to attain desired network modularity, and we can reproduce the spectral scaling trends associated with this topological class. Considering that each individual structure has a different spatial distribution of voids yet fall in the same topological class and the finding that BOO parameters, which are energy-like quantities, are largely

unaffected by this redistribution of voids hint that the globular protein landscape might be inherently degenerate at this level of coarse-graining.

A small positional disorder is needed to recover the necessary spread around the peaks that are present in the protein RDF, but these displacements do not disrupt other properties. Unlike simple liquids, the amount of positional disorder that needs to be introduced is rather limited. The combined RMC–SA procedure recovers the average RDF of 210 proteins by perturbing the on-lattice residues of a finite-sized structure with voids. The magnitude of the displacements produced by the reconstruction procedure does not significantly differ for the surface or core residues and stays well below the nearest neighbour distance ($\sim$3.8 Å) of proteins. Neither bond orientational order parameters (BOO) nor the network characteristics are affected by the RMC–SA simulations. All the resulting structures are shown to nontrivially approximate several topological and geometrical features of real proteins. Most notably, the established correspondence rules out icosahedral order as a relevant structural feature for proteins, in stark contrast to other amorphous systems such as dense liquids where icosahedral coordination is an inherent characteristic. We note that the node removal probabilities and the displacements can be biased to generate structures such that the surface nodes would be in a more loosely packed and positionally disordered state in comparison to the core nodes. As a part of future work, we will pursue this strategy as a possible route to establish a more refined correspondence between the artificial structures and the real proteins.

A comparison of vibrational properties through the ANM model shows that the static comparisons extend to the study of structural flexibility. We show that the generated models closely mimic the basic dynamical features of globular proteins by reproducing the spectral properties of the Hessian matrix. Therefore, the correspondence between the defect laden SC clusters and real residue networks is not superficial. Our approach is novel and uniquely general in the sense that it enables ensemble-vs.-ensemble comparisons without targeting individual protein structures.

We, however, note that the structure generation method does not ensure that the resulting structures form self-avoiding walks through nearest neighbour contacts. One possible strategy for ensuring self-avoiding walk property is to deepen the crystal lattice analogy to introduce extended defects. Most notably, a polycrystalline grain structure can be deployed wherein our ground rules could be imposed to generate smaller fragments. Such smaller self-avoiding fragments generated on each grain would then be tied together by reorienting each grain to yield the final self-avoiding polymer. However, to realize this, we need a set of well-defined rules to determine the optimal size of the grains and a meaningful texture map (orientation distributions of the grains), which we currently lack.

Finally, there has been a sustained interest in the computational manipulation of targeted self-assembly [84, 85] and building simple pair potentials that produce various open structures with controlled defect concentration as their ground state [86, 87]. Such a directed self-assembly approach is not readily applicable to proteins as it is hard to define candidate ground states with simple rules. However, culling novel tools from inverse optimization efforts might be a viable alternative to the widely used statistical potentials. The current work, by singling out SC arrangement as a candidate background and presenting a physically motivated recipe for structure generation, is a step forward in this direction.

## Supplementary data

Supplementary data are available at *COMNET* online.

## Acknowledgements

The authors acknowledge computational support provided by the Rensselaer Center for Computational Innovations. All figures, except for Fig. 4 (see figure caption for relevant references), were prepared with the Matplotlib graphics environment in Python [88].

## Funding

The National Science Foundation (NSF) (1538730 and 1825254), in part; and the Scientific and Technological Research Council of Turkey (TUBITAK) (117F389), in part.

## References

1. Branden, C. & Tooze, J. (1999) *Introduction to Protein Structure*, 2nd edn. New York, NY: Garland Science.
2. Petsko, G. A. & Ringe, D. (2004) *Protein Structure and Function*. London: New Science.
3. Atilgan, A. R., Durell, S. R., Jernigan, R. L., C., D. M., Keskin, O. & Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.
4. Tozzini, V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, **15**, 144–150.
5. Bahar, I. & Rader, A. J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–592.
6. Atilgan, C., Okan, O. B. & Atilgan, A. R. (2010) How orientational order governs collectivity of folded proteins. *Proteins*, **78**, 3363–3375.
7. Atilgan, C., Okan, O. B. & Atilgan, A. R. (2012) Network-based models as tools hinting at nonevident protein functionality. *Annu. Rev. Biophys.*, **41**, 205–225.
8. Saunders, M. G. & Voth, G. A. (2013) Coarse-graining methods for computational biology. *Annu. Rev. Biophys.*, **42**, 73–93.
9. Sinitskiy, A. V. & Voth, G. A. (2013) Coarse-graining of proteins based on elastic network models. *Chem. Phys.*, **422**, 165–174.
10. Go, N. & Taketomi, H. (1978) Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA*, **75**, 559–563.
11. Go, N. (1983) Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, **12**, 183–210.
12. Takada, S. (1999) Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA*, **96**, 11698–11700.
13. Hills, R. D., JR & Brooks, C. L., 3rd (2009) Insights from coarse-grained Go models for protein folding and dynamics. *Int. J. Mol. Sci.*, **10**, 889–905.
14. Sadoc, J.-F. & Mosseri, R. (1999) *Geometrical Frustration*. Cambridge: Cambridge University Press.
15. Glotzer, S. C. & Solomon, M. J. (2007) Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.*, **6**, 557–562.
16. Torquato, S. & Stillinger, F. H. (2010) Jammed hard-particle packings: from Kepler to Bernal and beyond. *Rev. Mod. Phys.*, **82**, 2633–2672.
17. Mannige, R. V. & Brooks, C. L., III (2010) Periodic table of virus capsids: implications for natural selection and design. *PLoS One*, **5**, e9423.
18. De Graef, M. & McHenry, M. E. (2012) *Structure of Materials: An Introduction to Crystallography, Diffraction and Symmetry*. Cambridge: Cambridge University Press.
19. Torquato, S. (2018) Perspective: basic understanding of condensed phases of matter via packing models. *J. Chem. Phys.*, **149**, 020901.
20. Chothia, C. (1975) Structural invariants in protein folding. *Nature*, **254**, 304–308.
21. Dill, K. A. (1990) Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
22. Onuchic, J. N. & Wolynes, P. G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.
23. Tribello, G. A., Ceriotti, M. & Parrinello, M. (2012) Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, **109**, 5196–5201.
24. Soyer, A., Chomilier, J., Mornon, J.-P., Jullien, R. & Sadoc, J.-F. (2000) Voronoï tessellation reveals the condensed matter character of folded proteins. *Phys. Rev. Lett.*, **85**, 3532.
25. Liang, J. & Dill, K. A. (2001) Are proteins well-packed? *Biophys. J.*, **81**, 751–766.
26. Shen, M., Davis, F. P. & Sali, A. (2005) The optimal size of a globular protein domain: a simple sphere-packing model. *Chem. Phys. Lett.*, **405**, 224–228.

**27.** Hummer, G., Garde, S., García, A. E., Paulaitis, M. E. & Pratt, L. R. (1998) The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Natl. Acad. Sci. USA*, **95**, 1552–1555.

**28.** Hillson, N., Onuchic, J. N. & García, A. E. (1999) Pressure-induced protein-folding/unfolding kinetics. *Proc. Natl. Acad. Sci. USA*, **96**, 14848–14853.

**29.** Roche, J., Caro, J. A., Norberto, D. R., Barthe, P., Roumestand, C., Schlessman, J. L., Garcia, A. E., García-Moreno E., B. & Royer, C. A. (2012) Cavities determine the pressure unfolding of proteins. *Proc. Natl. Acad. Sci. USA*, **109**, 6945–6950.

**30.** Prigozhin, M. B., Liu, Y., Wirth, A. J., Kapoor, S., Winter, R., Schulten, K. & Gruebele, M. (2013) Misplaced helix slows down ultrafast pressure-jump protein folding. *Proc. Natl. Acad. Sci. USA*, **110**, 8087–8092.

**31.** Rashin, A. A., Iofin, M. & Honig, B. (1986) Internal cavities and buried waters in globular proteins. *Biochemistry*, **25**, 3619–3625.

**32.** Hubbard, S. J., Gross, K.-H. & Argos, P. (1994) Intramolecular cavities in globular proteins. *Prot. Eng. Des. Sel.*, **7**, 613–626.

**33.** Graziano, G. (2007) Cavity size distribution in the interior of globular proteins. *Chem. Phys. Lett.*, **434**, 316–319.

**34.** Gaines, J. C., Smith, W. W., Regan, L. & O'Hern, C. S. (2016) Random close packing in protein cores. *Phys. Rev. E*, **93**, 032415.

**35.** Gaines, J. C., Clark, A. H., Regan, L. & O'Hern, C. S. (2017) Packing in protein cores. *J. Phys. Condens. Matter*, **29**, 293001.

**36.** Godzik, A., Kolinski, A. & Skolnick, J. (1993) Lattice representations of globular proteins: how good are they? *J. Comput. Chem.*, **14**, 1194–1202.

**37.** Lau, K. F. & Dill, K. A. (1989) A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, **22**, 3986–3997.

**38.** Bechini, A. (2013) On the characterization and software implementation of general protein lattice models. *PLoS One*, **8**, e59504.

**39.** Hart, W. E. & Newman, A. (2006) Protein structure prediction with lattice models. *Handbook of Computational Molecular Biology* (S. Aluru, ed.). Boca Raton, FL: Chapman & Hall/CRC, pp. 1–24.

**40.** Covell, D. G. & Jernigan, R. L. (1990) Conformations of folded proteins in restricted spaces. *Biochemistry*, **29**, 3287–3294.

**41.** Hinds, D. A. & Levitt, M. (1992) A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA*, **89**, 2536–2540.

**42.** Li, H., Helling, R., Tang, C. & Wingreen, N. (1996) Emergence of preferred structures in a simple model of protein folding. *Science*, **273**, 666–669.

**43.** Coluzza, I., Muller, H. G. & Frenkel, D. (2003) Designing refoldable model molecules. *Phys. Rev. E Stat Nonlin Soft Matter Phys*, **68**, 046703.

**44.** Abeln, S. & Frenkel, D. (2008) Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, **4**, e1000241.

**45.** Abeln, S., Vendruscolo, M., Dobson, C. M. & Frenkel, D. (2014) A simple lattice model that captures protein folding, aggregation and amyloid formation. *PLoS One*, **9**, e85185.

**46.** Watts, D. J. & Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

**47.** Estrada, E. (2010) Universality in protein residue networks. *Biophys. J.*, **98**, 890–900.

**48.** McGreevy, R. L. & Pusztai, L. (1988) Reverse Monte Carlo simulation: a new technique for the determination of disordered structures. *Mol. Simul.*, **1**, 359–367.

**49.** McGreevy, R. L. & Howe, M. A. (1992) RMC: modeling disordered structures. *Annu. Rev. Mater. Sci.*, **22**, 217–242.

**50.** Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

**51.** Inui, T., Tanabe, Y. & Onodera, Y. (1990) *Group Theory and Its Applications in Physics*. Heidelberg: Springer.

**52.** Buchete, N.-V., Straub, J. E. & Thirumalai, D. (2003) Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J. Chem. Phys.*, **118**, 7658–7671.

**53.** Buchete, N.-V., Straub, J. E. & Thirumalai, D. (2004) Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. *J. Mol. Graph. Model.*, **22**, 441–450.

**54.** Meydan, C. & Sezerman, O. U. Representation of protein secondary structure using bond-orientational order parameters (T. Shibuya, H. Kashima, J. Sese & S. Ahmad, eds) *Pattern Recognition in Bioinformatics*. Heidelberg, Germany: Springer, pp. 188–197.

**55.** Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. (1983) Bond-orientational order in liquids and glasses. *Phys. Rev. B*, **28**, 784–805.

**56.** Goshen, S., Mukamel, D. & Shtrikman, S. (1971) Application of the Landau theory of phase transitions to liquids-liquid crystals transitions. *Solid State Commun.*, **9**, 649–652.

**57.** Jarie, M. V. (1986) Landau theory of long-range orientational order. *Nucl. Phys. B*, **265**, 647–670.

**58.** da Silveira, C. H., Pires, D. E. V., Minardi, R. C., Ribeiro, C., Veloso, C. J. M., Lopes, J. C. D., Meira Jr, W., Neshich, G., Ramos, C. H. I., Habesch, R. & Santoro, M. M. (2009) Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, **74**, 727–743.

**59.** Pires, D. E. V., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M. & Meira, W. (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12**, S12.

**60.** Atilgan, A. R., Akan, P. & Baysal, C. (2004) Small-world communication of residues and significance for protein dynamics. *Biophys. J.*, **86**, 85–91.

**61.** Bahar, I., Atilgan, A. R. & Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173–181.

**62.** Turgut, D., Atilgan, A. R. & Atilgan, C. (2010) Assortative mixing in close-packed spatial networks. *PLoS One*, **5**, e15551.

**63.** Yilmaz, L. S. & Atilgan, A. R. (2000) Identifying the adaptive mechanism in globular proteins: fluctuations in densely packed regions manipulate flexible parts. *J. Chem. Phys.*, **113**, 4454–4464.

**64.** Demirel, M. C., Atilgan, A. R., Bahar, I., Jernigan, R. L. & Erman, B. (1998) Identification of kinetically hot residues in proteins. *Protein Sci.*, **7**, 2522–2532.

**65.** Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R. & Covell, D. G. (1999) Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.*, **285**, 1023–1037.

**66.** Bahar, I., Atilgan, A. R., Demirel, M. C. & Erman, B. (1998) Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.*, **80**, 2733–2736.

**67.** Plaxco, K. W., Simons, K. T. & Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.

**68.** Godsil, C. & Royle, G. (2001) *Algebraic Graph Theory*, vol. 207. New York, NY: Springer.

**69.** Newman, M. (2010) *Networks: An Introduction*. New York, NY: Oxford University Press.

**70.** Chung, F. R. K. (1997) *Spectral Graph Theory*. Providence, RI: American Mathematical Society.

**71.** Banerjee, A. & Jost, J. (2008) On the spectrum of the normalized graph Laplacian. *Linear Algebra Appl.*, **428**, 3015–3022.

**72.** Estrada, E. (2007) Topological structural classes of complex networks. *Phys. Rev. E*, **75**, 016103.

**73.** Fariselli, P. & Casadio, R. (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng. Des. Sel.*, **12**, 15–21.

**74.** Meyer, C. D. (2000) *Matrix Analysis and Applied Linear Algebra*. Philadelphia: Society of Industrial and Applied Mathematics.

**75.** Raghunathan, G. & Jernigan, R. L. (1997) Ideal architecture of residue packing and its observation in protein structures. *Protein Sci.*, **6**, 2072–2083.

**76.** Atilgan, A. R. & Atilgan, C. (2012) Local motifs in proteins combine to generate global functional moves. *Brief. Funct. Genomics*, **11**, 479–488.

**77.** Lorenz, C. D. & Ziff, R. M. (1998) Precise determination of the bond percolation thresholds and finite-size scaling corrections for the sc, fcc, and bcc lattices. *Phys. Rev. E*, **57**, 230–236.

**78.** HUMPHREY, W., DALKE, A. & SCHULTEN, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 27–38.
**79.** MOMMA, K. & IZUMI, F. (2008) VESTA: a three-dimensional visualization system for electronic and structural analysis. *J. Appl. Crystallogr.*, **41**, 653–658.
**80.** TURGUT, D. (2011) Network characterization of packing architecture for condensed matter systems. *Ph.D.*, Sabanci University, Ystanbul, Turkey.
**81.** ELBER, R. & KARPLUS, M. (1986) Low-frequency modes in proteins: use of the effective-medium approximation to interpret the fractal dimension observed in electron-spin relaxation measurements. *Phys. Rev. Lett.*, **56**, 394–397.
**82.** HALILOGLU, T., BAHAR, I. & ERMAN, B. (1997) Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, **79**, 3090–3093.
**83.** LINDGREN, B. W. (1993) *Statistical Theory*, 4th edn. Boca Raton: Chapman & Hall/CRC.
**84.** COHN, H. & KUMAR, A. (2009) Algorithmic design of self-assembling structures. *Proc. Natl. Acad. Sci. USA*, **106**, 9570–9575.
**85.** SHERMAN, Z. M., HOWARD, M. P., LINDQUIST, B. A., JADRICH, R. B. & TRUSKETT, T. M. (2020) Inverse methods for design of soft materials. *J. Chem. Phys.*, **152**, 140902.
**86.** MARCOTTE, É., STILLINGER, F. H. & TORQUATO, S. (2011) Unusual ground states via monotonic convex pair potentials. *J. Chem. Phys.*, **134**, 164105.
**87.** BATTEN, R. D., HUSE, D. A., STILLINGER, F. H. & TORQUATO, S. (2011) Novel ground-state crystals with controlled vacancy concentrations: from kagomé to honeycomb to stripes. *Soft Matter*, **7**, 6194–6204.
**88.** HUNTER, J. D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.