Distribution of volume, microvoid percolation, and packing density in globular proteins

Jenny Farmer,1,2 Sheridan B. Green,3 and Donald J. Jacobs1,4

1Department of Physics and Optical Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
2Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
3Department of Physics, Yale University, New Haven, CT 06520, USA
4Center for Biomedical Engineering and Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

(Dated: October 23, 2018)

A fast and accurate grid-based method with low memory requirement is presented to calculate volume characteristics in molecular systems. The distribution of volume and packing density is characterized in globular proteins, where void space is decomposed into microvoid volume and cavities based on a spherical test probe with variable radius. A scan over test probe radius is mapped onto a site percolation problem for microvoid volume. Finite-size scaling is applied to determine critical exponents, which are found to be consistent with connectivity percolation exponents in three dimensions. Disparate results in the literature regarding packing density in the core of a protein compared to on its surface, and with respect to protein size, is elucidated in terms of microvoid volume within a unified implicit-solvent model. By parameterizing the model to match the results of explicit-solvent models that agree with experimental data, we verify that packing density within globular proteins is spatially uniform and independent of protein size.

PACS numbers: 87.14.Ee, 64.60.Fr

Keywords: atomic packing, packing density, protein cavities, percolation theory, finite-size scaling

I. INTRODUCTION

The volume of space occupied by and surrounding a protein is an important property that affects protein stability and function. In particular, globular proteins are generally densely packed in the native state, yet ramified interior void spaces are present due to imperfect packing. Local volume information is important for pressure unfolding, as well as structure prediction and drug design. Boundary volume, related to accessible surface area, correlates with hydrophobic transfer energies, which are a driving force in conformational changes. Furthermore, clefts and pockets that form on a protein surface are distinguishing characteristics that are effectual in predicting catalytic areas and binding sites. Volume decomposition strategies frequently employ a spherical test probe for purposes of classification, where the results depend on the test probe radius. For example, a cavity defines a region in space that is large enough to hold a probe of a certain size. As probe size changes, different types of volume contributions merge or separate. It is common to use a probe radius of 1.4 Å to represent a water molecule in order to infer functional mechanisms from local volume characteristics. Identifying how cavities change in size within a protein for different probe sizes is useful in quantifying how various types of local volumes cluster as probe radius is continuously varied as an independent variable.

Many computational methods have been implemented to calculate protein volume, void space, surface area, as well as to identify clefts and channels. Grid-based methods have been shown to have a competitive advantage over analytical methods for large macromolecular structures. In this paper, we report a novel grid-based algorithm with several distinctive features that provide important advantages over existing methods. First, we incorporate compressive elasticity into the test probe as a means to identify cavities and to mitigate artifacts from the discrete lattice. Second, scalability and performance are optimized using a generalized cluster labeling technique, similar to the original Hoshen-Kopelman (HK) method, to characterize volume in a protein as a multi-component percolation problem. Third, a variable-sized spherical test probe is utilized to classify clusters of different volume types, where a direct connection to percolation theory is made by mapping the probe to a site percolation probability. Fourth, we model the effect of

* djacobs1@uncc.edu
protein dynamics by averaging results over a narrow distribution of test probe sizes. Finally, after partitioning the volume into distinct types, we assign contributions to each volume type on a per atom basis, enabling the accurate calculation of packing density.

A survey of the literature shows that a variety of methods that calculate local packing density yield inconsistent results, although most methods report at least marginally higher packing in the core. The nature of packing density remains an open problem, in large part due to inconsistent operational definitions. Conceptually, packing density is a measure of the percentage of the protein volume that is comprised of the atoms themselves, not including the empty space between the atoms. Voronoi tessellation should provide an accurate method for computing packing density. Unfortunately, the Voronoi method has difficulties near the surface of a protein when the solvent is not explicitly modeled. In past works, these technical problems were avoided by limiting the analysis to buried residues.

Attempts to extend applicability to surface residues have been handled in a variety of ways, such as by imposing boundary conditions to model a solvation shell surrounding the protein, by excluding certain intractable surface volumes from the calculations, or by strategically placing water molecules around the protein. A different approach, called occluded surface packing (OSP), computes packing density by extending lines from each atom perpendicular to its atomic surface, until the lines either intersect with another atomic surface or reach a length equal to the diameter of a water molecule. The lengths of these lines are used to determine the packing density, which is typically less than the corresponding Voronoi packing estimate. Yet another approach, by Liang and Dill, applies a separate definition for surface packing density that considers pockets along the protein surface.

The various implicit solvent models differ in their predictions of local packing density for residues that are exposed to solvent. A straightforward way to resolve this problem is to perform a molecular dynamics (MD) simulation of a protein with explicit solvent at the all-atom level and apply Voronoi methods. Notwithstanding subtle technical problems of its own, this direct approach arguably provides the most accurate representation of the microenvironment that surrounds a protein but suffers from being too computationally expensive to use in high-throughput applications in structural bioinformatics and drug design. Our implicit model is parameterized to match the results from these all-atom MD simulations.

This paper is outlined as follows. In Section II, we define four distinct, mutually exclusive types of volume and describe the algorithm for assigning each grid point’s volume type and cluster connectedness. In Section III, we benchmark the accuracy and speed of the method and present results for volume characteristics over a dataset of 108 globular proteins that are diverse in size and fold structure. Microvoid clusters are analyzed in the context of percolation theory. By employing finite size scaling, we show universal behavior in cluster size scaling for microvoid volume across a diverse set of globular proteins. In Section IV, we quantify packing density characteristics in terms of partial volumes. Local packing density depends on how much of a residue is exposed to solvent and how partial volume is partitioned within an implicit solvent model. Within our unified framework, we elucidate how different trends appear depending on the operational definition of partial volume. We conclude in Section V by highlighting the advantages of our method and discussing future work and applications.

II. METHOD

A. Definitions of volume types

Different definitions of volume types are possible because the decomposition of protein volume is not unique. For example, solvent-accessible volume and solvent-excluded volume are two commonly invoked terms with varying definitions throughout the literature. Molecular volume and accessible volume are commonly defined by the prescription given by Richards, which involves using a spherical probe of radius that rolls around the entire surface of a protein. The contact surface and accessible surface, first calculated by Lee and Richards, are two distinct areas carved out by this process. The contact surface is also called the molecular surface, or Connolly surface, and is defined by the leading edge of the spherical probe. The accessible surface is obtained by extending the contact surface at each point by the probe radius. The molecular volume is the volume enclosed by the contact surface and the accessible volume is enclosed by the accessible surface.

In this work, molecular volume is expressed as a sum of microvoid volume, cavity volume, and vdW volume. Any space within the molecular volume that is not occupied by atoms (i.e., not part of the vdW volume) is called void space, which we divide into cavity and microvoid volume. Cavity volume consists of regions that are large enough to accommodate the spherical probe, and we denote the remaining volume as microvoid. A schematic of the volume decomposition is shown in Fig. Note that the minimum span of any cross section of microvoid volume must be less than the probe diameter. Therefore, clusters of contiguous microvoid volume are highly ramified because large pockets of microvoid volume cannot exist. Microvoid volume, also called "dead space", typically percolates throughout a protein. To our knowledge, this microvoid volume, which represents the vast majority of molecular volume, was not previously calculated as a quantity of interest. Note that, unlike the vdW volume, the clustering of microvoid and cavity volumes will depend on probe size.

In addition to the molecular volume, we include an ex-
FIG. 1: Volume space definitions. The vdW volume of the protein atoms is denoted by the cross-hatched circles. The white space represents microvoid, and a single cavity is shown in black. The gray surrounding layer is the solvent extended volume layer, with a default depth of two times the diameter of the probe molecule.

tended volume layer beyond the molecular surface. The surface of this extended volume layer extends past the contact surface up to length $L_s$, forming an outer shell surrounding the protein. All grid points that are located within the extended surface are included in the volume calculation. Setting $L_s = 4R$ allows for two layers of probes to fit inside the extended volume layer, which is useful for studying solvation properties of residues by encapsulating the solvation shell. Setting $L_s = R$ recovers the accessible volume defined by Richards. We also parameterize $L_s$ by comparing to all-atom MD simulation results that indicate uniform packing density throughout a protein. In this case, a function $L_s(R)$ is constructed to ensure that the average packing density on the surface of a protein is the same as within its core.

We define boundary volume as the final volume type. Boundary volume consists of the volume in the extended layer as well as all cavity volume that has a connecting path to the extended volume layer. For example, a cylindrical channel through a protein would be part of boundary volume provided that a spherical probe can traverse through channel. However, if the probe size is too large to enter the channel on both ends, then the channel forms a separate microvoid region. When the probe radius reflects the size of a water molecule, boundary volume identifies the pathways where water can penetrate into the protein. In summary, we define four volume types: vdW, cavity, microvoid, and boundary. Adding these four partitions reconstitutes the molecular volume augmented by the extended volume surrounding a protein.

B. Volume connectedness

We employ a modified three-dimensional version of the Hoshen-Kopelman (HK) method. The HK method is widely used to identify which points on a grid form connected clusters, utilizing the union-find algorithm that is common to many computer science applications. The HK method is an efficient cluster-labeling technique that assigns a common label to a set of neighboring grid points. The time complexity is linear in the number of grid points and the method requires only a single scan through the lattice to identify each distinct cluster. We map an all-atom protein structure onto a cubic lattice with lattice spacing $a$. The lattice is traversed by processing successive pairs of planes of grid points along their common normal direction. The leading plane is denoted $F$ for front and the trailing plane is denoted $B$ for back, where $B$ lags $F$ by one lattice constant in the dimension normal to the planes. The size of these planes is set such that the edges of the $B$- and $F$-planes reside just outside the extended volume. The initial $B$-plane is placed just outside of the extended volume of the protein such that all grid points in this plane are initialized as boundary volume. Similarly, all grid points on the edges of each plane are initialized as boundary volume. The first cluster label is used to represent boundary volume. We note that there can only be one boundary volume cluster based on our definition and boundary conditions. No other initialization is required. After the HK scan is complete, the protein fits within a box with side lengths $\{L_x, L_y, L_z\}$. All grid points on each face of this box share the first cluster label. The HK clustering label technique merges neighboring grid points with the same volume type into clusters with unique cluster labels. Note that every grid point inside this box is assigned to one of the four volume types defined above. In the case of boundary volume regions, a point is discarded if it is outside the extended volume that is set by the $L_s$ distance cutoff.

During the HK scan, each grid point within the $F$-plane is systematically evaluated row by row. First, the volume type of a grid point must be identified. If a grid point is within the vdW radius of an atom, it is identified as vdW volume. Otherwise, it is marked as void space. To further distinguish between cavity and microvoid volume requires a compressive energy test applied to the test probe, which is described in the next subsection. Each grid point is identified as one of three volume types: vdW, microvoid or cavity. Note that boundary volume represents a specific cluster of grid points that are marked
as cavity, which is automatically identified based on the boundary conditions described above. Second, an unla-
beled grid point that is in question within the $F$-plane is assigned a cluster label based on its neighboring grid points where volume types have already been identified within the $F$-plane and $B$-plane. In particular, the mini-
imum cluster label among all its neighboring grid points with the same volume type is assigned to this unlabeled grid point, followed by the merging of cluster labels ac-
cording to the HK method. Otherwise, if no previously-
scanned neighbors of the same volume type are present, a new cluster label is introduced and the grid point in question is assigned to this cluster. Third, when the $F$-
plane is completely scanned, no further history of the previously scanned $B$-plane is needed. The $F$-plane then serves as the new $B$-plane and the current $B$-plane is recycled as the new $F$-plane in an alternating fashion.

The HK cluster labeling technique identifies contigu-
ous space defined by connected grid points of the same volume type. Two grid points of the same volume type are connected if they are nearest, second-nearest or third-
nearest neighbors on the cubic lattice. Altogether, a grid point can be connected to up to 26 neighbors. We consid-
ered models with lower connectivity by limiting connec-
tions between grid points to be nearest neighbors only or nearest and second-nearest neighbors only. All the models showed qualitatively similar results. However, the greater connectivity model that we use in all of the results presented here was found to approximate contin-
uous space better based on this model obtaining lower uncertainties in volume predictions across random orienta-
tions of the protein when it is embedded on the cubic lattice.

During the HK scan, the size of each uniquely labeled cluster using a canonical label, is accumulated on the fly using counters, which can subsequently combine as differ-
ent clusters merge and canonical labels condense. Thus, the memory requirement to store grid points and their as-
associated characteristics is two-dimensional, where we de-
fine the $z$-axis to be normal to the $B$- and $F$-planes such that $L_z L_y$ gives the cross-sectional area of the planes. After the HK scan is complete, the volume for each in-
dividual cluster is recorded using histograms for cluster statistics. To minimize memory requirements, we can mini-
imize the cross-sectional area by selecting a special direc-
tion for the $z$-axis that coincides with the eigenvector of the largest eigenvalue of the protein’s moment of inertia tensor. Reducing memory also increases speed when grid points no longer need to be processed. How-
ever, there is roughly an average 10% reduction in mem-
ory and increase in speed when embedding a large set of diverse globular proteins on the lattice, each optimized to have an orientation with minimum cross-sectional area. Although a particular $z$-axis direction can provide greater runtime and memory savings for long, cigar-shaped pro-
teins, in this work all of our results are based on the averaging of over 300 uniformly random orientations per protein calculation.

In tandem with tracking cluster volumes for all void types, we calculate partial volumes by assigning the vol-
ume and type of each grid cell to its nearest-neighbor atom. As each grid point is classified, it is recorded by volume type in cumulative counters for the atom whose vdW shell is closest to that grid point. Thus, if the dis-
tance between a grid point $g$ and atom $a$ from center to center is $r_{ga}$, the distance that the grid point is away from the atom’s vdW shell is $\delta_{ga} = r_{ga} - R_{vdW}$, where $R_{vdW}$ is the vdW radius of the atom. As such, each atom is ascribed four partial volumes (vdW, cavity, mi-
crovoid, and boundary) to describe its immediate local environment. The total vdW volume is non-additive be-
cause the spherical balls representing vdW interactions generally overlap among atom pairs that are covalently bonded. Note that the vdW volume is weakly dependent on conformation whereas cavity, microvoid and boundary volume all strongly depend on the protein conformation. For example, partial boundary volume is a measure of how exposed an atom is to solvent. The partial bound-
ary volume of a residue is obtained by summing all partial boundary volumes from all atoms within the residue. At the residue level, partial boundary volume correlates well with solvent accessible surface area (SASA) when com-
pared to DSSP and NACCESS [59, 60]. Partial volumes (of various types) can also be assigned to the backbone or sidechain portions of residues, where this distinction has been shown to be critically important for implicit solvation models [61].

C. Mitigation of discrete lattice artifacts

Classifying void space requires checking if a test probe can fit into a region of space without overlapping with any neighboring atom’s vdW shell. If the boundary of a test probe centered at a grid point does not overlap with the boundary of any atom, then all grid points that are en-
closed by the test probe are classified as cavity. Im-
portantly, a grid point can be part of cavity while being arbitrarily close to the boundary of an atom. Clearly, it is not sufficient to check if a grid point is greater than a minimum distance from all nearby atoms to classify it as cavity or microvoid. To take advantage of the efficiency in the HK method where only the $B$- and $F$-planes are stored in memory at any given time, each grid point that appears in the $F$-plane must be identified on the fly as micro-
void or cavity. Therefore, a procedure is needed that can displace the location of the test probe (defined by its geometric center) from the initial grid point position to a position that no longer overlaps with neighboring atoms while keeping the initial grid point within the span of the test probe. We employ a compressive energy test that is capable of performing this procedure.

The compressive energy test is constructed through a pseudo-potential energy function of the form
where the length $b = R + R_{vdW}$, with $R$ the probe radius and $R_{vdW}$ the atom type’s vdW radius. This pairwise potential energy is calculated for all atoms that create a compressive force on the test probe. The same spring constant is used for all atom types. In addition, we apply a geometric constraint that requires $r_{pg} < R$, where $r_{pg}$ is the distance between the test probe and the grid point in question. The model parameters are adjusted to maintain distance tolerances to within 0.0125 Å, 0.3 Å, and 0.1 Å grid sizes, the relative errors are relative error can be fixed at respectively 2.5%, 4.2%, and 12.5%. Alternatively, the relative geometric constraint that requires a compressive force on the test probe. The same spring to microscopic interactions between the atoms reduces related to grid size. Said another way, uncertainties due to microscopic interactions between the atoms reduces the need for exact geometric results. Therefore, we set a maximum energy threshold of $E_{\text{max}} = \frac{1}{2}k(0.0125 \text{ Å})^2$. Since units are arbitrary, we set $k = 2$.

Determining if a grid point is microvoid or cavity requires numerical relaxation to minimize the compressive energy on the test probe of radius $R$ that is initially placed at a reference grid point. This minimized energy is denoted as $E(R, n|k)$, where $n$ is the number of atoms within the local environment surrounding the reference grid point that can possibly clash with the probe and $k$ is the number of iterations applied in the minimization process before termination. Note that $n$ counts all atoms that could potentially clash with the test probe, but the actual number of clashes, if any, will be far less than $n$. This is because as the test probe moves in a certain direction, clashes are more likely to occur with distant atoms positioned in that direction, while distant atoms positioned in the opposite direction will become too far away to create a compressive force. Operationally, $n$ gives the array size needed to account for all possible clashes the test probe could encounter as it moves away from the reference grid point. If $E(R, n|k) > E_{\text{max}}$, then the grid point is counted as microvoid; otherwise, the point contributes to cavity because the probe does not overlap with any atom’s vdW radius by more than the distance threshold. A schematic for compressive elasticity is shown in Fig. 2.

The compressive energy minimization process is summarized by five key properties. First, a proof by construction shows that a test probe cannot clash with $n \leq 3$. When a test probe has compressive energy, it typically clashes with several atoms. Consequently, when $0 < E(R, n|k) < E_{\text{max}}$, the overlap errors per atom are typically much lower than the distance tolerance. Second, it is only possible to misidentify a true cavity region as microvoid but not vice versa. This misidentification can occur if the relaxation method fails to locate the true energy minimum, denoted as $E(R, n|\infty)$. Specifically, $E(R, n|\infty) < E_{\text{max}} < E(R, n|k)$ is the erroneous case where the grid point is a true cavity but is classified as a microvoid. Third, we employ conservative convergence thresholds to abort the energy minimization process when the rate of energy decrease becomes slow while the gap in energy above $E_{\text{max}}$ remains high. That is, once $E(R, n|k) - E(R, n|k + 1) \ll E(R, n|k + 1) - E_{\text{max}}$, we terminate the energy minimization. The error cases are quite rare and only occur when convergence is slow and $E(R, n|\infty) \lesssim E_{\text{max}}$. Fourth, a larger test probe will result in more grid points being identified as microvoid because more neighboring atoms will typically contribute to compressive energy. Additional compressive springs slow down the energy minimization process. Fifth, when a grid point is identified as microvoid using a small probe, it will remain a microvoid for any larger probe because $E(R_2, n_2|\infty) \geq E(R_1, n_1|\infty)$ when $R_2 > R_1$ where it is also the case that $n_2 \geq n_1$.

Taking these properties into account leads to an algorithm that, for each grid point, begins by minimizing the compressive energy for a small probe size and iteratively considers larger sizes until the probe size of interest is reached. We find that a schedule for probe sizes that is linear in the volume of the probe works best. However, we determine the number of different probe sizes in the schedule to be $N_{\text{probe}} = \lceil \frac{R}{0.5\text{ Å}} \rceil$, where $R$ is the probe radius of interest. For example, if we wish to consider a 4 Å probe, we employ a total of 8 different probes whose volumes are equally spaced. As we iterate over the probe sizes, if the energy relaxation clearly shows strain energy, then the grid point must be microvoid and the remaining, larger probe sizes need not be checked. However, if a grid point cannot be identified as microvoid due to lack of convergence or is identified as cavity for a probe size iteration, then the algorithm continues on to the next probe size. This process is repeated until the probe size of interest is reached. Note that misclassification rates are reduced by using a higher energy threshold for smaller probes in the loop. On the final probe size, more energy minimization iterations are allotted to reach convergence if necessary. Otherwise, the energy minimization procedure is the same.

In summary, an intrinsic problem to iterative relaxation methods is to decide upon a convergence threshold. An incorrect identification as microvoid indicates that the system had not yet reached its minimum energy when a decision was made and $E(R, n|\infty) < E_{\text{max}}$. But to maintain speed, some errors can be tolerated to avoid large number of iterations. Therefore, we tested the method by repeatedly making predictions on the same set of proteins with the same orientations while varying the maximum number of iterations, threshold criteria,
FIG. 2: Visual demonstration of the compressive elasticity model. Top: Compressed springs cannot be fully relaxed and maintain nonzero strain energy. Therefore, the initial grid point is classified as microvoid. Bottom: After relaxing the compressive energy for a smaller probe, the lack of remaining strain energy implies that the probe is able to fit within the available space. Thus, the initial grid point is classified as cavity.

and other parameters that alter the rate of convergence and performance. The most accurate combinations of parameters result in the algorithm becoming rather slow. While evaluating different parameterizations, we located discrepancy cases and checked the reasons for and the severity of the incorrect assignment of the grid point as microvoid volume. We then adjusted the model parameters to arrive at a low non-systematic error rate that is less than one incorrect microvoid assignment per $2.5 \times 10^5$ microvoid assignments. Overfitting is not an issue due to the model only having a handful of parameters compared to hundreds of millions of generic local molecular geometries that were tested. Note that no error occurs when a grid point is assigned as cavity. In addition, because we average the calculation over 300 random rotations and the errors virtually never occur at the exact same point in the protein, the overall error rate is less than $10^{-6}$ percent in misidentification. A much more substantial source of error is due to the finite resolution of the lattice when discretizing space, which we demonstrate is also acceptably low in Section III.

In this work, we calculate volume characteristics of proteins based on their static structure determined via X-ray crystallography. Although these static structures are commonly employed to characterize the native state of a protein, there are ambient vibrations about the equilibrium conformation. To account for local volume fluctuations as atoms vibrate, we consider a variable test probe size. This requires us to calculate the average volume characteristics of a protein over an ensemble of test probe sizes. These probe sizes are drawn from a Gaussian distribution with a specified mean radius and a standard deviation set to 5% of this mean radius. With negligible additional computational cost, we randomly rotate the protein about its center of mass before embedding it onto the lattice in tandem with performing the calculation over the ensemble of test probe radii. We find that averaging over a narrow spectrum of probe radii helps to smooth out the discrete nature of the data quite well. By modeling probe sizes with slight variations and using the compressive energy model to capture off-lattice structural variations, we effectively mitigate artifacts arising from geometrical idealizations based on static structure. As a consequence, accurate calculations of volume can be obtained for any grid size provided it is equal to or less than the mean probe radius.

III. VOLUME CHARACTERISTICS

To test the volume calculations, we apply our method to a subset of structures selected from the Top 500 dataset of globular proteins, originally published by Hobohm and Sander [62]. This list was compiled to ensure good structure quality, resolution, and non-redundancy amongst members of the set. We removed
all structures from the original Top 500 with any missing chains, residues, or atoms. Additionally, we removed structures with an incomplete biological unit definition and those with fewer than 50 residues. The remaining 108 structures span a range of sizes from 50 to over 800 residues, providing a solid basis for testing and benchmarking the method as well as establishing general trends in volume characteristics across globular proteins.

A. Performance and convergence properties

For each protein, 300 rotations in uniformly random directions for lattice embedding were performed for grid sizes 0.1–1.0 Å in linear increments of 0.1 Å and up to 76 logarithmically spaced probe radii ranging from the grid size to 4.0 Å. Mean values and variances over the ensemble of rotations were calculated for each volume type as a global quantity as well as for partial volumes per atom and residue. The CPU times were clocked on the same high-performance cluster using a single core with an Intel Xeon 2.93 GHz processor for all calculations. Fig. 3 summarizes how CPU times vary as a function of grid size, probe radius, and protein length. The results shown in Fig. 3a are for the case of varying grid size with a fixed probe radius of 1.4 Å. We find that $T_{CPU} \propto a^{-2.95}$, indicating that the method is linear in the number of grid points. The time complexity dependence on test probe radius is more complicated. Fig. 3b shows compute times versus test probe radius with a fixed lattice constant of $a = 0.1$ Å. The nearly linear relationship between $R$ and $\log(T_{CPU})$ demonstrates an approximately exponential time complexity for $R$. In both cases, CPU times for each of the 108 proteins were averaged and normalized by the protein’s sequence length (e.g. number of residues $N_R$).

A linear scaling of CPU time with protein length is seen in Fig. 3c which shows that the method is well suited for large molecular systems. Points are plotted for all 108 proteins for a grid size of 0.1 Å and probe radius of 1.4 Å. Thus, the empirically determined time complexity for processing a protein is given by $T_{CPU} \sim N_R e^{c/R}$, where $c$ is a constant.

In Fig. 4a and 4b, the coefficient of variation, defined as the ratio of the standard deviation to the mean and expressed as a percentage, is plotted as a function of grid size for various types of volumes. Figure 4a shows the convergence for each volume type with 300 rotations for a fixed probe size of 1.4 Å. The coefficient of variation decreases as grid resolution increases, demonstrating virtually negligible error with rotation, even for relatively large grid sizes. The convergence of the same quantities using the probe-averaging method with a mean radius of 1.4 Å is shown in Fig. 4b. The grid size dependence for the vdW coefficient of variation is essentially the same regardless of fixed or variable probe radius, confirming that vdW volume is independent of probe size. Interestingly, microvoid and boundary volumes have nearly constant coefficients of variation across all grid sizes, indicating that the random variation on the test probe radius introduces much more variance than the resolution effects. The reason for placing variation on the test probe radius is to better model soft vdW interactions and dynamic fluctuations in structure. Since probe size variation is the dominant cause for variations in our ensemble averaging, consideration of multiple rotations is not necessary. Nevertheless, randomizing over different protein orientations on a grid contributes negligible computational cost when performing an ensemble average over probe size.

The variation of cavity volume over random protein orientations for both fixed and varied probe size is shown in Fig. 4c. As with microvoid and boundary volumes, for fixed probe size, the variations in cavity volume decrease with decreasing grid size. However, the coefficient of variation is considerably larger than that of other volume types, which is explained by noting that cavity volume is small, yet they undergo relatively large percent changes in volume due to sensitivity in connecting to channels that extend to the protein surface. For a small probe radius, the channel opens up, allowing solvent to penetrate into the cavity. In this case, cavity volume becomes boundary volume. Similarly, two small microvoid clusters may merge together to form a cavity. There is also a greater degree of sensitivity to protein orientation when a probe radius is close to the critical threshold for a cavity to dramatically change size. These effects are enhanced by varying the probe radius, as shown in Fig. 4c. Again, the variation in probe radius increases the coefficient of variation for cavity volume and appears to dominate over the dependence on grid size.

Taken together, the plots in Figs. 3 and 4 provide little justification for using grid sizes much smaller than the probe size of interest. The much greater computational cost yields very little gains in accuracy due to intrinsic uncertainties from ambient dynamical fluctuations. Since the probe radius is usually taken to be 1.4 Å when considering volume characteristics in proteins, a grid size of 0.5 Å provides ample accuracy with calculation times that are competitive with the fastest methods reported in the literature. It is worth noting that methods such as alpha shape are limited to a static view of the protein unless averaging over multiple frames from a MD simulation. If our method were applied to an MD trajectory, we would use a fixed probe radius and not randomize the protein orientation on the grid since the trajectory naturally captures the dynamical fluctuations.

B. Linear scaling with protein length

We observe that some quantities depend strongly on protein length. In Fig. 5 the results are exemplified for the case that grid size and probe radius are 0.1 Å and 1.4 Å, respectively. Note that qualitatively similar correlations are found (not shown) when we replace the number of residues (i.e. protein length) with the number
of atoms. In Fig. 5a, we find that vdW volume and microvoid volume scale linearly with protein length. In contrast, cavity volume, shown in Fig. 5b, has only a slight tendency to increase with protein length. These trends have been previously established for cavity and vdW volumes using a variety of methods and datasets [1, 19]. However, to the best of our knowledge, microvoid as a separate quantity has not yet been considered. Figs. 5c and 5d show that the total number of distinct microvoid clusters scales linearly with protein length, and the number of distinct cavities (e.g. cavity volume clusters) also scales approximately linearly. Although it is intuitive to expect partial volumes to be extensive, like total volume, it is clear that cavity volume is not an extensive quantity, and as noted above is considerably smaller than other volume types. In particular, microvoid volume dominates the total void space. This is because cavity formation is thermodynamically unfavorable [4, 63, 64]. Hence, the presence of a cavity is expected to be of biological importance [14, 65–67]. Interestingly, the number of clusters for microvoid and cavity appear to be extensive. These results show that a significant number of globular proteins tend to have greater cavity volumes than a simple linear trend line based on extensive properties alone, which suggests we are observing evolution-driven effects captured in our dataset.

C. Volumes as a function of probe size

The classifications of cavity, microvoid, and boundary volume are dependent upon the size of the probe. As the probe size approaches zero, the probe can access virtually all space outside of the van der Waals radii of the protein atoms. Thus, in this limit, cavity and microvoid volumes shrink to zero and the volume is decomposed into either vdW or boundary volume. As the size of the probe increases, crevices become unreachable by the probe external to the molecular surface, thus creating void space. The probe-size dependence of vdW, microvoid, and cavity volumes expressed as fractions of the molecular volume is shown in Fig. 6a. All of these volumes are averaged over 300 rotations for 108 proteins with error bars that are no larger than the symbols representing the volume type. For small probe size, the molecular volume is nearly equal to the vdW volume, whereas for large probe size, the relative fraction of vdW volume decreases while microvoid volume contributes much more to the molecular volume. Note that the vdW volume divided by the molecular volume is a common definition for packing density and the results we obtain in Fig. 6a agree with earlier studies [1, 2, 24].

The cavity volume, on the other hand, accounts for a
FIG. 5: (a) Characteristics as a function of protein length, averaged over 300 rotations for grid size 0.1 and probe radius 1.4 for (a) total vdW and microvoid volume, (b) total cavity volume, (c) number of microvoid clusters, and (d) number of cavity clusters.

tiny fraction of the molecular volume for most proteins in our dataset. The peak in cavity volume that occurs around 0.8 Å, as shown in Fig. 5a, may potentially be a universal characteristic for globular proteins. This result does not depend on grid size provided it is less than 0.8 Å to capture this resolution. To study this further, a magnified view of the cavity volume versus probe radius is plotted for each of the 108 proteins as separate curves in Fig. 6b. In addition, we overlay the probability density [68] for the probe radii at maximum cavity volume for each of the 108 proteins. The observed noteworthy consistency in the probe radius where the maximum cavity volume occurs across proteins is quite interesting and we posit the following explanation. Because globular proteins tend to pack well, the size of void space within local atomic environments is determined mainly by vdW interactions. A small probe radius allows the boundary volume to more easily penetrate the protein, substantially reducing cavity volume. A large probe radius creates more microvoid volume, which also substantially reduces cavity volume. A small range in probe size naturally appears where the cavity volume should be maximized. However, this geometrical and physical explanation does not appear to have a biological interpretation. It is worth mentioning that we observe a slight shift in the peak probe size that maximizes cavity volume when using alternative vdW radii definitions, as discussed below.

Since thermodynamic forces tend to limit cavity formation, the presence of a large cavity is a signature of an important functional characteristic of a protein. We note that, for several proteins in our dataset, very large peaks in total cavity volume are found for probe radii above 2 Å (see Fig. 6b). For each of these cases, we identify a single large pocket within the protein that is connected to solvent accessible tunnels with a much smaller opening (e.g., a lollypop shape) that will prevent large molecules from entering. In particular, PDB:1QFM is a protein in the dataset that is not plotted individually on Fig. 6b for
visual clarity, as it contains a large cavity over 8000 Å³ large with an opening of approximately 4 Å in diameter that out-scales the axes. These results highlight the value of scanning proteins with multiple probe sizes as a means of gaining a more complete picture of potentially important functional characteristics. This also emphasizes the importance of the local volume characteristics that will be discussed in Section IV.

D. Microvoid percolation

Percolation theory provides a ubiquitous framework to understand clustering phenomena in a wide array of systems and has been previously used to characterize both protein volume and cavity volume within proteins. Here, we investigate the characteristics of microvoid percolation. Fig. 6a shows that microvoid volume is a monotonically increasing function of probe radius. Fig. 7 illustrates how the largest cluster of microvoid volume in a protein changes as probe radius increases. For large probe sizes, the largest cluster of microvoid volume will span the entire protein, forming a percolating cluster. Note that for visualization purposes, the coordinates of all grid points, their volume types, and cluster labels are saved in memory as the HK method scans over a protein to identify specific clusters. These coordinates and cluster labels are saved such that they can be visualized using PyMOL. For example, in Fig. 7, the protein volume is shown in light gray with black shading representing the largest microvoid cluster for probe sizes below, at, and just above the percolation threshold. The largest cluster at the threshold in Fig. 7b has an irregular shape and does not span the protein in all directions simultaneously. Microvoid clusters are typically found to be non-spherical and they decrease in sphericity (as defined in) with increasing microvoid cluster size.

To characterize microvoid volume in the context of percolation theory, we set $L_s$ to a fixed value to maintain a constant system volume as probe size is varied. At each probe size we run the calculation over 300 random orientations. We define $V_m$, $V_c$ and $V_b$ to be the microvoid, cavity, and boundary volumes, respectively. Probe size is then mapped to site probability using the relationship:

$$p = \frac{V_m}{V_m + V_c + V_b},$$  \hspace{1cm} (2)$$

Note that vdW volume is omitted from consideration in Eq. (2) because it can never be occupied by microvoid volume. The denominator in Eq. (2) represents total void space, which is a fixed value as probe size changes when $L_s$ is held constant. The mapping of this site probability to probe size for all proteins is shown in Fig. 8a for the case that $L_s = 0.55$ Å. We consider different values for $L_s$ ranging from 0 to 5.6 Å, finding that the site percolation threshold monotonically decreases as a function of $L_s$. This is obvious because $V_b$ increases with $L_s$, which lowers the range of possible site probabilities. Furthermore, we consider a range of grid sizes from 0.1 Å to 0.3 Å. Although we arrive at the same conclusion regarding percolation properties for all choices of $L_s$ and grid size considered, the analysis becomes more difficult to perform using very small or large values of $L_s$. The choice of $L_s = 0.55$ Å provides a good working range of site probabilities for all proteins in our dataset.

For finite systems, an operational definition for the effective site percolation threshold $p_c$ is the site probability associated with the peak value of the reduced second moment (RSM) which is defined as:
FIG. 7: Visualization of the largest microvoid cluster, shown in black, (a) below, (b) at, and (c) just above the percolation threshold.

\[
\text{RSM} = \sum_s s^2 n_s \sum_s n_s,
\]

where \( s \) is the size of a microvoid cluster taken as number of grid points, \( n_s \) gives the number of clusters of size \( s \), and the summation runs over non-percolating clusters only. The percolation threshold for each of the 108 proteins is calculated individually and is markedly similar for all proteins. To see this, the dark circles in Fig. Sa mark the probe size and site probability at the percolation threshold for each individual protein. Fig. Sb shows RSM as a function of site probability using the collective cluster size statistics across all 300 rotations. Similarly, Fig. Sc plots the fraction of percolated realizations as a function of the site probability for each protein by using the peaks in RSMs that are calculated separately for each rotation. In almost all cases, the expected sigmoid-shaped curve describing a single sharp transition is present.

Among the 108 proteins, only two exceptions were found that have multiple discrete transition jumps leading to plateaus in the percolation probability curves, shown as black solid lines in Fig. Sc. Interestingly, the two exceptions are the largest protein in our dataset (839 residues) and one of the smallest proteins (61 residues). A third exception, shown as the rightmost dashed black line in Fig. Sc is an example of a non-spherical protein with an unusually shallow percolation probability curve with a larger threshold value compared to typical cases. The large amount of variation in RSM characteristics, illustrated in Fig. Sd, makes it challenging to apply percolation theory to microvoid volume because proteins of similar size are inherently inhomogeneous. Nonetheless, the site probabilities and percolation thresholds are remarkably similar across a diverse collection of globular proteins.

All of the results discussed thus far in Figs. Sa, Sb, and Sc were produced using a lattice constant of 0.3 Å. Fig. Sd shows the probability density for the percolation thresholds for three different lattice constants, which highlights a few interesting characteristics. A shift in \( p_c \) towards smaller values occurs with decreasing grid size, while the spread in the distributions narrow. This qualitative trend is expected due to the greater resolution of spatial connectivity as the grid size decreases. Quantitatively, it is curious that the coefficient of variation for all three cases is approximately constant. Note that grid sizes greater than 0.4 Å cannot be used to determine the microvoid percolation threshold because the grid spacing cannot be greater than the probe radius, otherwise the probe size would be smaller than the grid resolution. However, according to standard percolation theory, the values for critical thresholds do not have any significance regarding scaling laws.

The remaining analysis regarding microvoid percolation uses a grid size of 0.3 Å, which yields a mean percolation threshold of \( p_c = 0.114 \) over the 108 proteins. Notice that this percolation threshold for site percolation is significantly smaller than that for the random cubic lattice, which is approximately 0.311 \([76, 81]\). Unlike a cubic lattice that has homogeneous site probability throughout the system, in proteins, the grid points representing vDW volume and the outer limits of the boundary volume are not accessible to microvoid. It is therefore clear that for large \( L_s \), the microvoid volume will be a small percentage of total void volume, trivially decreasing the percolation threshold. More interestingly, the geometrical constraints that originate from the underlying molecular structure induce correlation between microvoid volume and protein conformation. In a general context, correlated site probabilities have been shown to be an important factor that can lower the percolation threshold compared to an uncorrelated random lattice \([82, 85]\).

In a standard percolation problem, a system can be defined by a simulation box of linear dimension \( L \), where it is possible to numerically estimate \( p_c(L) \) with small statistical errors. The \( p_c(L) \) estimates for different system sizes can be used subsequently in the context of finite-size scaling (FSS). In comparison, the application of percolation theory to microvoid volume within a protein is un-
conventional in three respects. First, the site percolation probability given in Eq. (2) is correlated due to the covalently bonded molecular structure underlying the native conformation of the protein being analyzed. Second, protein structure is not homogeneous and it is therefore not possible to increase system size of a given protein by simply doubling its size, for example. Third, FSS involves combining different size proteins together, but the statistics per protein is very limited. In addition, our sample size of 108 proteins is not especially large in the context of percolation theory. Below, we present an analysis of microvoid percolation using FSS to extract critical exponents. When considering the quality and size of our dataset, the deviations that we find in the critical exponents relative to connectivity percolation in three dimensions are not sufficiently large to claim microvoid percolation lies within a different universality class. Said another way, the critical exponents that describe microvoid percolation in globular proteins are consistent with the standard connectivity percolation critical exponents in 3 dimensions.

We apply FSS to estimate a hypothetical percolation threshold corresponding to the limit of an infinite-sized globular protein (i.e., not repeat proteins). The threshold in this theoretical limit can be extracted from the effective probability $p_c(L)$, determined by the peak in the RSM, where $L$ gives the linear dimension of the system. We calculate the linear size $L$ of a protein as done previously in the literature [1, 86], where $L$ is defined as the average of the protein’s maximum extent along each coordinate axes:

$$L = \frac{1}{6} \sum_{j=1}^{3} (x_{j,max} - x_{j,min}).$$  (4)

The most extreme coordinates of all the atoms in each
protein are used in each of the x, y, and z directions. We checked different variants of linear size with respect to rotation, such as by using the original orientation of the protein from the input PDB files, the average of this quantity over 300 random rotations, and the greatest value from the 300 random rotations. These variants have no effect on the subsequent analyses and conclusions. The results shown below are based on Eq. (4) with the simplest implementation, namely with calculating $L$ from the initial orientation of the protein from the input PDB file. According to FSS, the percolation threshold for a finite-size system has the following form:

$$p_c(L) = aL^{-\frac{d}{\nu}} + p_c(\infty),$$

where $p_c(L)$ is empirically determined from the RSM, $a$ is a fitting parameter and $p_c(\infty)$ is the percolation threshold for the hypothetical infinite-size globular protein. The finite-size scaling exponent $\nu$ is universal in the sense that it depends only on the dimension of the system.

The critical exponent $\beta$ can be determined by calculating the strength of the infinite cluster according to the equation

$$P_\infty \sim |p - p_c|^{\beta},$$

where $P_\infty$ is the probability that a random site is a member of the infinite cluster in a hypothetical infinitely large protein. Applying FSS to Eq. (5) with $p_c(\infty)$ yields a scaling law [87] of the form

$$P_\infty(L|p = p_c(\infty)) \sim L^{-\frac{d}{\nu}},$$

where $P_\infty$ is evaluated at the same site probability, equal to $p_c(\infty)$, for all systems regardless of their size. The exponent $\frac{d}{\nu}$ for this power law is calculated from the slope of the straight line that results when log($P_\infty(L)$) versus log($L$) for each protein at the estimated $p_c(\infty)$ (data not shown). This means the estimate for $\frac{d}{\nu}$ depends on the extrapolated value of $p_c(\infty)$. Unfortunately, it is difficult to accurately extract $p_c(\infty)$ from Eq. (5) due to the noise in the data. We therefore develop a systematic procedure to extract the critical exponents and $p_c(\infty)$. This procedure will work for any percolation problem and, to our knowledge, it is a novel approach.

The first step is to treat $p_c(\infty)$ as an independent variable that we call $z$. Note that a conservative range of $z$ is considered such that $z_{\min} \leq p_c(\infty) \leq z_{\max}$ is true based on Eq. (5) and associated data, $p_c(L)$. We then determine the $\frac{d}{\nu}$ exponent from Eq. (7) as explained above for different values of $z$ that represent the putative $p_c(\infty)$ within the range $[z_{\min}, z_{\max}]$. After this scan is complete, we have $\frac{d}{\nu}$ as a function of $z$. Furthermore, other exponents can also be obtained based on universal scaling relationships. For example, the fractal dimension $d_f$ and critical exponent for cluster statistics $\tau$ can be calculated through the universal scaling relations

$$d_f = d - \frac{\beta}{\nu} \quad \text{and} \quad \tau = 1 + \frac{d}{d - \frac{\beta}{\nu}},$$

where $d_f$ and $\tau$ are treated as dependent variables and the dimension of space $d = 3$. Fig. 9a shows the fitted values of $\frac{d}{\nu}$ over a range of $z$ and from the relationships in Eq. (5), $d_f$ and $\tau$ are plotted on the same graph as a function of $z$.

As a second step, we estimate the fractal dimension. Fig. 9b shows the volume of the largest microvoid cluster in a protein at the critical threshold (based on the peak in the RSM) as a function of the cluster’s length, $\ell$. The volume of this largest cluster is expected to follow a power-law scaling that gives the fractal dimension $d_f$ according to

$$V_{\text{max}} = b\ell^{d_f}.$$
The next step is to apply Eq. (5) to this last equation in similar fashion as was done for $P_\infty$. In addition, cluster statistics are collected at the same $p_c(\infty)$ for all system sizes as the data from 108 proteins is combined. This combined data is described by

$$n_s(L|p = p_c(\infty)) \sim L^{-\frac{d}{\nu} + \frac{f}{\nu}}$$

where it must be that $f(x) \to x^{-\tau}$ for $x \gg 1$. This form recovers the usual cluster size scaling when $L \to \infty$. It appears this scaling form will allow the cluster statistics to collapse onto a single curve for more than one set of exponents. This freedom allows us to ignore the $\sigma$ exponent dependence.

As shown in Fig. 9c, we scale the cluster statistics data by introducing the scale factor in the $x$-direction as $L^{-\frac{1}{\nu}}$ and in the $y$-direction as $L^{\frac{1}{\nu}}$ to create a data collapse. The values of $\tau$ and $\nu$ used in these scaling factors come directly from the procedure above and are held fixed. Note that the quality of the data collapse shown in Fig. 9d is similar to any other data collapse we obtained by allowing the scaling factors to be optimized by treating $\tau$ and $\nu$ as free parameters. The characteristic leveling off at the large cluster size tail reflects poor sampling in the largest cluster sizes, with only one or two per bin [73, 88, 89].

Within the scaling region, the slope of the resulting data-collapsed curve in a log-log plot should be $-\tau$, which we use as an independent check on the overall consistency. To get this slope, the cluster statistics are scaled and then binned logarithmically. The circles in Fig. 9d represent the binned averages and the gray dots include all of the raw data. The straight line shown has a slope of $-2.040 \pm 0.380$, which encloses the accepted value of 2.19 for 3D connectivity percolation as well as the value estimated above in the third step ($\tau = 2.134 \pm 0.033$). Taken together, the exponents reported here are derived self-consistently and agree fairly well with the connectivity percolation exponents in 3D considering the inhomoge-
We use Bondi radii for all subsequent discussion.

IV. PACKING DENSITY CHARACTERISTICS

The long-standing accepted value of the mean packing fraction in the core region of globular proteins is approximately 0.75. In a recent study focused on a re-evaluation of vdW radii, a mean packing density of 0.56 within core regions of globular proteins was reported. In our work, the mean packing density in core regions agrees with the accepted value of 0.75 when Bondi vdW radii are employed and 0.56 is reproduced when the vdW radii proposed by Gaines et al. is used. Clearly, this discrepancy is related to the modeling of the vdW interactions, which reinforces our view that the results presented above are more physically meaningful when vdW interactions are modeled as soft interactions in the presence of dynamical fluctuations. We use Bondi radii for all subsequent discussion.

For Voronoi methods, the local packing density $\phi$ is generally defined as

$$\phi = \frac{V_{vdW}}{V_{cell}},$$  

(10)

where $V_{cell}$ is the Voronoi volume of a group of atoms of interest and $V_{vdW}$ is the total vdW volume for all atoms contained within $V_{cell}$. The packing density given by Eq. (10) can be applied at different spatial resolutions that range from a single atom, to a residue, as well as up the level of all protein atoms to determine a mean packing density for the entire protein. Voronoi methods make calculating the packing density simple in principle, but as discussed in Section I, the details of how to obtain the Voronoi volume of atoms on the surface of a protein within an implicit solvent model critically affects the accuracy of the density calculations. In a recent study focused on obtaining accurate packing densities in proteins, water molecules were treated explicitly to obtain a realistic solvent boundary at the protein surface, MD simulations were used to account for atomic fluctuations, and the Voronoi method was improved through empirically derived weighting parameters to accurately partition space among atoms. It is worth mentioning that using weight factors in the Voronoi method is critical for a realistic tessellation when dealing with differing atom sizes (i.e., vdW radii) found in proteins. Within this explicit-solvent framework, the volumes associated with residues buried in the core of a globular protein are approximately the same as those on the surface. Moreover, this result was shown to be consistent with experimental data. Thus, the implication is that packing density is indeed approximately a constant throughout a globular protein in its native state. Previous results suggested proteins are packed more densely in the core compared to the surface.

Considering the work of Esque et al., we elucidate why differences in predictions for local packing density occur across various models. Our partial volume calculations provide a means for studying intrinsic packing densities of core and surface residues separately. We propose an initial definition for packing density in terms of partial volumes as

$$\phi_0 = \frac{V_{vdW}}{V_{vdW} + V_m + V_c}.$$  

(11)

This frequently invoked definition calculates packing density as the ratio of vdW volume over molecular volume. Notice that partial boundary volume is not included in $\phi_0$. However, the partial boundary volume, as defined when $L_s = 5.6 \, \text{Å}$, is used to rank order how deeply buried a residue is within a protein according to the buried fraction $f_b$, given as

$$f_b = \frac{V_m}{V_m + V_b}.$$  

(12)

Based on Eq. (12), it can be seen that completely buried residues have no boundary volume and hence $f_b = 1$, whereas $f_b \to 0$ for residues at the surface when $V_b \gg V_m$. All residues are ranked by $f_b$ for each protein and split into two equal groups based on the median $f_b$ of that protein. Reference to core or surface residues are those that are respectively above or below the median $f_b$ for a given protein. In this way, each protein contributes the same number of samples to both groups. We also classified residues into buried and exposed groups such that buried residues have no boundary volume, whereas the exposed residues have a nonzero amount of boundary volume (no matter how small).

Distributions for packing density among the two groups of residues that are classified as buried or on the surface using a probe radius of 1.4 Å and grid size of 0.5 Å are shown in Figs. 10a and 10b and 10c. In Fig. 10a, the packing density distribution among core residues agrees well with Voronoi and OSP estimates. The packing density of residues near the surface, however, is slightly higher (c.f. Fig. 10a), contrary to the predictions of many other methods. In particular, surface packing densities from OSP are predicted to be lower than core densities, dipping as low as 0.2 to 0.4. Voronoi methods estimate different surface packing densities depending on how the solvent is modeled at the protein surface, but typically the packing density of surface residues is found to be slightly lower than that in the core. To understand the reason for differences in packing density between the core and the surface of a protein, the definition for packing density must be generalized to included boundary volume.

An alternative definition for packing density is given as

$$\phi_1(R, L_s) = \frac{V_{vdW}}{V_{vdW} + V_m + V_c + V_b},$$  

(13)
where, for the moment, probe radius is $R = 1.4$ Å and $L_s = 5.6$ Å, as above. The distributions for the $\phi_1$ packing density for buried and surface residues is shown in Fig. 10b. Not surprisingly, the most buried residues have the same average packing density compared with $\phi_0$ because these residues do not carry partial boundary volume at all. However, there is a decrease in packing density for partially buried residues and a dramatic decrease for surface residues. This large effect is due to the generous 5.6 Å layer of boundary surrounding the protein. The dashed line in Fig. 10b shows the distribution of packing density for exposed residues when $L_s$ is set to 2.8 Å, which gives a 1-water molecule boundary layer that is also employed in the OSP method. The criteria produce surface packing density ranges in good agreement with OSP calculations.

When modeling water molecules explicitly within a simulation box, no water molecules count as void space, although there is microvoid volume between water molecules. In an implicit solvent model, void space in the extended volume layer represents vdW, cavity and microvoid volumes lumped together. The question then becomes how much of this extended volume should be attributed to microvoid volume and subsequently assigned as partial volume to surface atoms. The answer can be obtained by generalizing the definition of packing density to include a free parameter $r_c$ to control the amount of boundary volume to retain:

$$
\phi(R, r_c) = \frac{V_{vdW}}{V_{vdW} + V_m + V_c + V_b(r_c)}.
$$

Specifically, we count only those boundary grid points that extend beyond the molecular surface up to a cutoff distance of $r_c$. Note that Eq. 11 and Eq. 13 are special cases of Eq. 14. We then find the optimal $r_c$.
Packing Density

shown in Fig. 10c, for a probe radius of 1.4 Å, a cut-off for core residues is equal to that for surface residues. As to satisfy the condition that the average packing density for core residues is equal to that for surface residues. As shown in Fig. 10c for a probe radius of 1.4 Å, a cut-off distance of 0.35 Å results in uniform packing density throughout globular proteins, matching quantitatively to the finely-tuned Voronoi tessellation method \[40\].

It is clear from Eq. (10) that if an explicit water model is employed, packing density will be independent of probe radius. To achieve uniform packing density that is independent of probe size, it is of interest to determine the \( r_c \) that keeps packing density uniform as a function of probe radius \( R \). The dependence of \( r_c \) on \( R \) is shown in Fig. 10c. This curve exhibits a maximum. Interestingly, the maximum cutoff distance occurs near the probe size consistent with a water molecule. Additionally, this maximum is relatively flat, meaning a cutoff distance of 0.35 Å is sufficient to yield roughly uniform packing density over a range of probe sizes between 1.3 Å to 2.1 Å. The lower cutoff values for smaller probe sizes is somewhat intuitive as we expect less microvoid to be present when a smaller probe is used. The reason for a decreasing cutoff for larger probes is less obvious but suggests a geometric relationship between the relative sizes of the protein residues and the solvent molecules at the surface. It is worth noting that similar results were obtained using the vdw radii of Gaines et al. \[72\]. We also replaced all the vdw radii of different atom types with a single value and varied this single value to remove possible frustration effects caused by different atomic sizes not fitting regularly. Similarly-shaped cutoff curves were found in all cases, albeit with different \( r_c \) values and peak radii. This extended investigation of packing fractions finds no particular fine-tuned dependence on the size of water.

Protein length is another factor that affects packing density, but the correct dependence has not yet been determined in the literature \[11\] [24] [91] [94]. Applying Eq. (11) to the total protein volumes for all structures in our dataset, we find that packing density is smaller for larger proteins as shown in Fig. 11. If we include a generous portion of boundary volume as in Eq. (13), the length dependence reverses, in agreement with OSP. Finally, by applying Eq. (14) with a cutoff of 0.35 Å, we find spatially-uniform packing densities that are independent of protein length. Thus, our unified model with a single adjustable parameter can be tuned to explain the differences in packing density between various models that have been reported in the literature.

V. DISCUSSION

We implement a fast and accurate grid-based method with a low memory requirement for calculating the volume properties of molecular systems. The method is based on a generalization of the Hoshen-Kopelman cluster-labeling technique for identifying and uniquely labeling all contiguous clusters of the same volume type. This approach classifies microvoid, cavity, and boundary volume as separate quantities. Microvoid represents void space that is too small to accommodate a spherical probe of a specified radius. Grid points are identified on the fly as either cavity or microvoid by employing a compressive energy model that mitigates discrete lattice effects by allowing a tolerance for atomic clashing. Furthermore, a narrow distribution of different probe sizes is sampled in order to account for dynamical atomic motion when static structures are analyzed.

A representative sample of globular proteins has been selected to test and benchmark this method as well as to collect statistics on typical properties. The clustering of microvoid volume is analyzed using percolation theory. We show that microvoid clustering is well described by standard connectivity percolation. Furthermore, partial volumes for each volume type are assigned to each atom to give detailed maps of how volume and packing density are distributed throughout a protein. Our implicit-solvent model for packing density provides a unified framework with a single adjustable parameter to control the amount of microvoid volume outside the molecular volume of a protein that is attributed to the local volume of a protein atom. After parameterizing the model by matching to the results from explicit-solvent models that are consistent with experimental data, the model predicts packing density to be uniform within globular proteins, independent of the protein size. By leaving the adjustable parameter for packing density free, we recover the predictions of previous implicit-solvent models that lead to disparate results, thus explaining from a unified perspective the discrepancies in predictions for a variety of packing density properties found in the literature.

As an implicit-solvent method, our approach has many applications to high-throughput analysis of protein structures. The solvation layer length and probe radius are
key parameters that can be adjusted when investigating the distribution of volume in proteins. For example, microvold volume characteristics have a universal character in the native state across diverse globular proteins. In ongoing work, the universal cluster statistics is being employed to place confidence levels on protein structure predictions to better identify decoys. Identifying cavities at different length scales as the test probe radius is varied also helps glean insight into protein function. In future applications in condensed matter systems with explicit solvent. Moreover, this method will generalize to any molecular system, and thus can support many applications in condensed matter systems with explicit modeling of all atoms.

ACKNOWLEDGMENTS

This work was partly supported by NIH grant GM101570. SBG was partly supported by the NOAA Ernest F. Hollings Scholarship during this work. The authors wish to thank Dennis Livesay for stimulating discussions in the early stages of this project.
