Article

Assessment of Globularity of Protein Structures via Minimum Volume Ellipsoids and Voxel-Based Atom Representation

Mateusz Banach

Department of Bioinformatics and Telemedicine, Faculty of Medicine, Jagiellonian University Medical College, Medyczna 7, 30-688 Kraków, Poland; mateusz.banach@uj.edu.pl

Abstract: A computer algorithm for assessment of globularity of protein structures is presented. By enclosing the input protein in a minimum volume ellipsoid (MVEE) and calculating a profile measuring how voxelized space within this shape (cubes on a uniform grid) is occupied by atoms, it is possible to estimate how well the molecule resembles a globule. For any protein to satisfy the proposed globularity criterion, its ellipsoid profile (EP) should first confirm that atoms adequately fill the ellipsoid's center. This property should then propagate towards the surface of the ellipsoid, although with diminishing importance. It is not required to compute the molecular surface. Globular status (full or partial) is assigned to proteins with values of their ellipsoid profiles, called here the ellipsoid indexes (EI), above certain levels. Due to structural outliers which may considerably distort the measurements, a companion method for their detection and reduction of their influence is also introduced. It is based on kernel density estimation and is shown to work well as an optional input preparation step for MVEE. Finally, the complete workflow is applied to over two thousand representatives of SCOP 2.08 domain superfamilies, surveying the landscape of tertiary structure of proteins from the Protein Data Bank.

Citation: Banach, M. Assessment of Globularity of Protein Structures via Minimum Volume Ellipsoids and Voxel-Based Atom Representation. *Crystals* **2021**, *11*, 1539. <https://doi.org/10.3390/cryst11121539>

Academic Editor: Abel Moreno

Received: 16 November 2021

Accepted: 7 December 2021

Published: 9 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: bioinformatics; bounding ellipsoid; computational geometry; convex hull; globular protein; hydrophobic core; kernel density; protein domain; quaternary structure; tertiary structure

1. Introduction

Steven Brenner and colleagues begin their 1995 paper with following statement [1]: “The structure of a protein can elucidate its function, in both general and specific terms, and its evolutionary history.” In the next paragraph, the same authors add: “(...) protein structures can be fundamentally understood in ways that most of their sequences cannot.” Indeed, both sequence and structure of proteins have their own means of providing insight into biological role and ancestry of these molecules [2]. This in turn allows them to be classified, stored in databases and explored in the light of current state of knowledge.

Finding homologous coding regions in the primary structure is done through (multiple) alignment, carried out by tools like the BLAST program suite [3–5]. By the end of 2021, the total number of protein sequences collected by the Universal Protein Resource (UniProt) [6,7] has surpassed 550,000 in the manually curated pool, supported by over 225,000,000 automatically annotated entries. On the other hand, the number of protein structures available in their largest public repository, the Protein Data Bank (PDB) [8–10], is currently “only” over 180,000, but grows every week.

Given two proteins—a query and a reference—if their sequence-based similarity does not provide enough information to ascertain the function of the first, following the second quote from above, structure-based comparison may be of help. This of course demands that computer models of these molecules are available, typically in form of entries in PDB. Such approach is viable due to the fact that homology at the structural level has higher conservation ratio than at the sequence [11]. Proteins from different organisms may reach the same fold in order to accomplish their biological activity. As a result, the number of

these folds is much lower than the number of sequences, permitting more direct human input in their analysis. In addition, tertiary and quaternary structure acts as a mirror in which a crucial factor—the influence of water environment on function, folding and complex formation—is reflected [12–15].

As all proteins share (to some extent) structural similarities with each other [16], there is a natural motivation for scientists to organize those relationships in databases, focusing on both their shape and evolutionary origin. SCOP (Structural Classification of Proteins) [17–19] and CATH (Class/Architecture/Topology/Homologous superfamily) [20–22] are two projects which maintain and expand such databases. The basic structural unit undergoing classification by them is the protein domain. Both projects employ a mixture of manual curation and computer-aided automated matching to keep up with continuous batches of new PDB contents.

Like CATH, SCOP distributes its domains on four levels of a hierarchy tree [16]. These levels are (from bottom to top): family—sequence-based clustering, superfamily—structure-based clustering, common fold—similar secondary structure composition sharing same topological arrangement, and class—fold grouping based on general secondary structure composition for user's convenience. SCOP tracks following classes: *a* (all alpha), *b* (all beta), *c* (alpha/beta), *d* (alpha+beta), *e* (multi-domain), *f* (membrane and cell surface), *g* (small), *h* (coiled coil), plus *i* (low-res), *j* (peptides), *k* (designed) and *l* (artifacts). As of version 2.08 (released in second half of 2021), the first eight SCOP classes constitute 1264 folds, 2134 superfamilies, 5170 families and 305,324 domains.

Protein domains are independent regions of the molecule, capable of having their own hydrophobic core [1]. Structures containing them subscribe to one of four main types: globular, fibrous, membrane and intrinsically disordered [23]. A globular protein is expected to have ellipsoidal shape and good solubility which should translate into a well-established hydrophobic core shielded from water by polar residues [24]. Such core is not found in fibrous proteins where all residues are exposed to the solvent. Membrane proteins inhabit specific environment of the cell wall which imposes on them specific shape and function that is lost when they are withdrawn from this scaffold [25]. Stability varies across “disprots” as their biological activity depends on global or regional flexibility [26]. Finally, a special mention is needed here for amyloids, undergoing both structural and hydrophobic core transformation from centric to linear [27,28]. These transformations can be—among other characteristics—tracked with the Fuzzy Oil Drop (FOD) model [29] (<http://fod.cm-uj.krakow.pl>).

FOD encloses the input structure in a 3D Gauss-based ellipsoidal capsule which simulates both the surrounding water environment and protein's idealized relation to it (absolute solubility). The resulting theoretical distribution of density of hydrophobicity is then confronted with observed (empirical) distribution of density of hydrophobicity originating from interaction between residues. Discrepancies between those two profiles allow functional analysis of the molecule in regard to its biological activity, core stability and election of putative protein–protein and protein–ligand interaction areas [30,31].

FOD is primarily interested in relationships in the data (*T* profile vs. *O* profile). It accepts its input structure regardless of whether it is globular or not, trying to fit it inside the 3D Gauss capsule. It does not need to verify that status—it is not its purpose—but knowing that status allows better understanding of the results and more informed data preparation. Such information may also provide useful for P-P contact area prediction or domain exploration and discovery, when one is looking for globular (or dense) regions in the structure, especially those that constitute segments of multiple chains [32].

In this paper we present a simple method for general assessment of shape of tertiary (and by extension—quaternary) structure of proteins for the purpose of determining how well they resemble a globule and to facilitate their comparison based on this property. By simple we mean one that is easily programmable in a generic-purpose language, easily pluggable as a subroutine for another program, has only one main parameter (which does not require tuning for specific molecules), works with every type of input (domain, chain,

complex, etc.) and does not need to rely on information provided by molecular surface solvers. Instead, we directly use voxels—cubes on a uniform lattice—to represent the protein's volume, which in our opinion (and based on obtained results) is an adequate substitute. Voxelization is not a novel concept in bioinformatics, often utilized for simplification of the molecular surface mesh with a variable level of detail [33], for example in ligand binding research [34]. But, to the best of our knowledge, the presented approach has not been tried yet in context of measurement of overall protein globularity, especially in combination with structural outlier detection based on kernel density.

Globularity is expressed here as the capability of atoms comprising the input structure to fill the smallest ellipsoid enclosing it (MVEE). Use of bounding ellipsoids in bioinformatics is also not new. Mentions of them can be found in the literature from 1980s [35] and they are employed in recent works [36,37]. Our approach is different though, going beyond the measurement of ellipsoid's semi-axes. However, we also do not aim for detailed mapping of the protein shape universe [36,38], but instead provide an instrument for investigation of one of its defining features.

We know that surface of proteins is not smooth [39]. Hence, our ellipsoid (the MVEE) should tightly fit to the input structure, but without being negatively influenced by protruding sidechains or other outliers such as disordered chain regions. At the same time, it must also not overdo it by leaving too much of the molecule's main body outside its surface. Next, the protein should be checked for significantly large cavities located towards the solvent and towards the interior. In other words, to pass as a globule, it should resemble a solid shape (a “drop”) rather than a loose chain. These conditions are verified by the presented algorithm, the Ellipsoid Profile (EP). “Profile” in its name refers to the distribution of measure of globularity (called the ellipsoid index) that moves from the center of the protein to the surface of its enclosing ellipsoid.

In Section 3, we first describe and demonstrate application of EP algorithm to several selected structures, displaying its possible input and output. After that we apply it to 2124 SCOP 2.08 superfamily representatives in order to present how it surveys the landscape of tertiary structure of proteins from the Protein Data Bank.

2. Materials and Methods

2.1. Data

Initial database of this work comprised 2065 domains—representatives of SCOP 2.08 superfamilies such as a.1.2: the alpha-helical ferredoxin. This information was obtained from the ASTRAL 2.08 compendium [18] downloaded from the SCOP website [40] (data available only for SCOP classes *a* through *g*).

We decided to omit the 7 multi-chain genetic domains (g1gk9.1, g1jmu.1, g1ko6.1, g1mso.1, g1n13.1, g1qtn.1, g1sse.1). We also replaced 43 domains found in structures marked as obsolete by PDB. In 34 cases the replacement PDB codes were supplied by PDB itself [41] and had matching domain entries already present in SCOP. PDB structures containing other 9 domains were either completely obsoleted or none of their superseding structures was currently available in SCOP. We replaced them with manually-selected members of the same superfamily: d1rlla3 → d1l9na2 (b.1.5), d1vtz0_ → d4oq9a_ (b.121.7), d2j0111 → d2zjru1 (d.325.1), d2j0181 → d2zjr31 (d.301.1), d2j01u1 → d2zjrn1 (a.144.2), d2j01v1 → d2zjro1 (b.155.1), d2wt3a3 → d2x41a3 (b.1.31), d2x1ya1 → d1r1ra1 (a.98.1), d4bkha1 → d5m33a1 (a.135.1). To minimize the differences, where applicable, we chose proteins from the same source organism. For completeness, we also included in the database representatives of superfamilies from SCOP class *h*: the coiled coil proteins. There is only 67 of those superfamilies with 7 folds in SCOP 2.08. Because class *h* is not included in ASTRAL, we used a simple selection system by taking first domain (sorted alphabetically) from each superfamily belonging to a protein not marked as obsolete by PDB. All structures from superfamily h.6.1 and their entire fold were discarded for this reason. Only one additional manual change was made here: d1xq8a_ → d2kkwa_ (h.7.1) in order to have in the database a higher-ranking structure of alpha-synuclein.

Above alterations increased the total number of input domains from 2065 to 2124 and the total number of unique PDB codes from 1850 to 1901. 1724 structures were obtained via X-ray diffraction, with mean resolution of 1.75 Å ($\sigma = 0.62$ Å). 149 had multiple models, 19 on average ($\sigma \approx 7$). In their case, only the first model was used.

Eight domains from the final database were chosen as examples for this paper. Their basic information is presented in Table 1. We refer to them in the text first by their names but later—for brevity—by their PDB codes. Six of those domains are sequence-wise equal with their parent chains (i.e., d1b79a_). Among other three, d1diva1 has sequence range of 56–149. Second domain from the same protein (d1diva2) spans residues 1–55 and its superfamily (d.100.1) is represented in ASTRAL records by d2hbaa1. Domain d2b1ya1 is equal with portion of chain A available in the PDB structure (missing residues 1–3), while domain d3bpd1 is shorter than chain A by one residue at each terminus (both classified by SCOP as artifacts, class I).

Table 1. Proteins used as examples. Data in Quaternary Structure column denotes composition of author-assigned biomolecule containing the domain shown in SCOP Domain column. Asterisk (*) marks complexes recreated with symmetry operators obtained from REMARK 350 records of the PDB header. Values after hash (#) correspond to number of solution NMR models present in the PDB file.

PDB Code	Molecule	Source Organism	Chain Length	SCOP Domain	Quaternary Structure	Refs.
1B79	DnaB Helicase	<i>Escherichia coli</i>	102 aa	d1b79a_ (a.81.1.1)	Monomer ($\times 4$)	[42]
1DIV	Ribosomal protein L9	<i>Bacillus stearothermophilus</i>	149 aa	d1diva1 (d.99.1.1)	Homo-2-mer *	[43]
1IIE	HLA-DR Invariant Chain	<i>Homo sapiens</i>	75 aa	d1iiea_ (a.109.1.1)	Homo-3-mer (#20)	[44]
1IS9	Endoglucanase A	<i>Clostridium thermocellum</i>	358 aa	d1is9a_ (a.102.1.2)	Monomer	[45]
1L2P	ATP Synthase B Chain	<i>Escherichia coli</i>	61 aa	d1l2pa_ (f.23.21.1)	Monomer	[46]
2B1Y	Hypothetical Protein Atu1913	<i>Agrobacterium tumefaciens</i>	101 aa	d2b1ya1 (b.156.1.1)	Homo-2-mer *	[47]
2KKW	Alpha-synuclein	<i>Homo sapiens</i>	140 aa	d2kkwa_ (h.7.1.1)	Monomer (#34)	[48]
3BPD	Uncharacterized Protein	<i>Archaeoglobus fulgidus</i>	91 aa	d3bpd1 (d.58.61.1)	Homo-7-mer ($\times 2$)	[49]
4B0H	dUTPase YncF	<i>Bacillus subtilis</i>	130 aa	d4b0ha_ (b.85.4.0)	Homo-3-mer	[50]

2.2. Algorithms

Below are brief descriptions of the four key algorithms utilized during research presented in this paper. Their common feature is that they all accept as their input a set of n points in d -dimensional space, which here represent atomic coordinates or their derivatives (effective atoms and voxels). These algorithms are also either parameter-free or their parameters tend to have universal defaults.

Leonid Khachiyan proposed an optimization algorithm for calculation of minimal volume enclosing ellipsoid (MVEE) [51] which received later improvements [52–54]. It produces a quadric surface form (positive-definite matrix M and center vector c) of an approximation of a smallest ellipsoid which encompasses the input set of points within error margin $\varepsilon > 0$. By calculating singular value decomposition (SVD) of M one obtains lengths of semi-axes of the output ellipsoid and its rotation matrix. This allows centering of the query at the origin and its alignment with axes of the coordinate system. Here we used Python port [55] of Nima Moshtagh’s implementation of MVEE in MATLAB [56].

Convex hull of a set of points is the smallest convex set of those points [57]. Since MVEE aims to surround its input data, when low error margin is requested ($\varepsilon \leq 0.01$, directing the algorithm to converge close to optimal surface rather than stop early with too many points still outside), it is beneficial to focus only on the input’s convex hull. By ignoring everything that should eventually lie inside the optimal ellipsoid, one obtains nearly the same output but—depending on distribution of the data—with a significant

calculation speed-up. It also offsets the higher number of iterations needed to produce such good approximation. Convex hulls can be calculated in 2D and 3D in $O(n \log n)$ time using—for example—the output-sensitive quickhull algorithm [58]. Here we used its qhull implementation [58] available through SciPy [59].

K-d tree (KDT) is a binary hierarchical data structure invented by Jon Bentley [60] for partitioning of space around input points into gradually smaller (hyper)rectangles. Once constructed, it allows $O(\log n)$ time answers to nearest neighbor search queries. These queries are often in form of looking for discrete number of nearest input points to an arbitrary point or all input points within specific distance to that point. Here we used the optimized k -d tree from SciPy [59] which implements the “sliding midpoint rule” after the work of Songrit Maneewongvatana and David Mount [61].

Kernel density estimation (KDE) [62,63], is a statistical method for non-parametric estimation of probability density function from a data sample. Compared to techniques based on histograms and k -nearest neighbors, it can produce a smoothed, continuous approximation when a suitable kernel function is employed [64]. Gaussian (normal) is one of the popular choices [65]. Input-wise KDE is the most time-consuming algorithm of the four, with $O(n^2)$ complexity. Here we used its implementation in SciPy [59] which uses Gaussian kernel and is capable of automated bandwidth selection.

2.3. Tools and Websites

3D images were rendered through PyVista [66], a streamlined Python interface to the Visualization Toolkit (VTK) [67]. Charts were plotted using Matplotlib library [68]. Results were obtained with the help of state-of-the-art open-source Python libraries for scientific computation [59,69]. Web applications of the Ellipsoid Profile algorithm and of the FOD model are available at <http://fod.cm-uj.krakow.pl> web server.

3. Results

Results are divided between five subsections. First four describe steps needed to implement the Ellipsoid Profile (EP) algorithm and explain rationale behind the specific approaches. Last Section 3.5 is dedicated to application of this algorithm to 2124 SCOP 2.08 superfamily representatives, surveying the landscape of globularity of protein domains present in structures currently available in PDB.

To reiterate the Introduction, we are aiming here for the best approximation of a smallest ellipsoid that encloses the input protein (MVEE) in a way that is less influenced by surficial details, but without erasing too much information. Then we check how well this ellipsoid is filled with atoms of the molecule using several synergistic metrics. This eventually allows us to classify it in terms of globularity and compare it with other proteins. A globular protein should in this sense have its atoms distributed mostly like a dense solid, with shape similar to a drop. It does not need to be entirely convex or perfectly filling the ellipsoid. Some distortions towards the solvent (ligand pockets, etc.) must be permitted, but its central part should be without large cavities. Structures which do not satisfy these requirements are considered non-globular by EP algorithm.

To achieve the above goals, we start with structure preparation (Section 3.1). Then we move to pre-MVEE step which involves selection of effective atoms and their separation into guides and outliers (Section 3.2). Next is ellipsoid fitting, molecule’s alignment and nearest voxel selection from uniform grid (Section 3.3). Presentation of EP algorithm concludes with calculation subroutine and description of its output, which comprises ellipsoid indexes and ellipsoid profile (Section 3.4).

3.1. Structure Preparation

Input to the EP algorithm has the form of computer model of a protein structure: a domain, chain, complex, several different molecules, etc.—generally any composition of residues and not necessarily amino acids. Artifacts, such as gaps in the sequence, missing atoms, or low resolution are also permitted. They distort the results but are not fatal.

The input structure must be prepared as usual. After loading its PDB file, the application should clean and normalize it by removing H₂O molecules, altlocs and applying MODRES records from the PDB header. The last part is highly recommended since atoms of modified residues may have no vdW radii specified (which is the case here), but can be otherwise well-approximated by their parents, i.e., selenomethionine by methionine. Furthermore, all procedures presented in this workflow ignore hydrogen by design. This means that atoms of this element can be removed at this stage to reduce the already short computation time even more.

3.2. Effective Atoms and Kernel Density

Error margin parameter of MVEE (ϵ) controls at which moment the optimization procedure is terminated. The lower the value of ϵ is, the more input points is contained in the resulting ellipsoid, which translates to better approximation of the optimal shape. Allowing some points to be left outside (but close to the surface) is preferred over large number of algorithm's iterations which might not yield significant improvements. A good balance seems to be reached with $\epsilon \approx 0.01$, which we use. On the other hand, higher values ($\epsilon > 0.1$) may cause MVEE solver to stop after only few steps, producing smaller ellipsoids still embedded in the input. This algorithm is also sensitive to the composition of input's convex hull. When applied to proteins, outliers (free loop regions protruding into the environment or long exposed sidechains) may cause it to output larger, emptier ellipsoids, which is undesirable when one is aiming to tightly contain the molecule in a representative way or is expecting similar results from similar structures.

In order to mitigate the negative effects mentioned above, we decided to calculate MVEE for effective atoms instead of actual atoms. The effective atom of a residue is understood here as the average position of all its heavy atoms. This means that missing sidechains can be managed without special cases and that positions of those sidechains should have lower impact on the output due to shift towards the backbone. It also greatly reduces the number of input points—down to the number of residues—roughly tenfold for a typical globular structure. Good approximation of the optimal ellipsoid can be now reached faster owing to its calculation only for the aforementioned convex hull, here of the effective atoms. It appears to yield another 50–90% decrease. For example, the highly globular Endoglucanase A (PDB code: 1IS9) has 2809 heavy atoms (5482 if hydrogen is also counted), represented by 358 effective atoms (12.7%) with 64 elements in their convex hull (17.9%, 2.3% total heavy atom ratio).

Use of effective atoms naturally causes some atoms to be left outside the ellipsoid, but we believe it provides better (smoother) approximation of the shape of the protein's body, with surficial noise pruned in a controlled way. This approach alone is however unable to prevent MVEE from being strongly influenced by more prominent outliers. dUTPase YncF (PDB code: 4B0H) is a perfect example of this phenomenon. In this trimer, residues at each C-terminus reach towards one of the sibling chains and form a short beta sheet with it (beta-clip). Enclosing any of those chains alone in an MVEE would leave a lot of empty space inside the ellipsoid, not satisfying the proposed globularity criterion (though it is satisfied by the complex where the monomers fuse together). A human observer would colloquially call them as “globules with a tail”, immediately noticing source of the problem. It is therefore desirable to be able to automatically detect such situations (at least partially) and present user with a choice or a warning. This is where kernel density estimation comes into use.

Calculation of Gaussian KDE for all effective atoms maps regions of their low and high density, denoted by low and high values of the approximated probability density function. In spatial sense, it measures the probability for given residue to be surrounded by large number other residues in respect to the whole structure. For simplicity we will call it now on as the density of (distribution of) effective atoms. We also scale it to [0, 1] range for easier visualization. We can do that because we are interested in density-based relationships instead of absolute values, which alone are meaningless.

Use of KDE has multiple benefits. First, it highlights residues with relatively too few neighbors, such as most of the “tail” protruding from the dense, globular body of 4B0H monomers (Figure 1a,d). However if low number of neighbors is the protein’s norm—like it is in ATP Synthase B Chain (PDB code: 1L2P), which is a simple, straight helix—only residues at the termini stand out (Figure 1b,e) after kernel’s bandwidth is automatically adjusted. When a few disjoint globular chains are the subject—four in DnaB Helicase (PDB code: 1B79)—high density is observed in their centers, as expected (Figure 1c,f). Again, the only effective atoms that fall here towards the lower value range of the approximated density function are those at chains’ surface.

Figure 1. Effective atom density profiles for: monomer of dUTPase YncF (4B0H) (a,d), ATP Synthase B Chain (1L2P) (b,e) and four monomers of DnaB Helicase (1B79) (c,f). Shaded regions on (d–f) mark residues with kernel density below threshold t (green squares under the red dashed line), calculated using three medians of the profile ($m = 3$). Their effective atoms—assumed to be structural outliers—are shown on (a–c) as green spheres. Guide effective atoms—non-green spheres on (a–c)—are colored in accordance with their kernel density values from (d–f). Thresholds matching first ($m = 1$) and second ($m = 2$) medians are symbolized by red dotted lines. Vertical dashed lines on (f) denote chain boundaries.

From Figure 1 it is clear that KDE can be used to—at least partially—detect outliers in the input structure: residues exhibiting the globally lowest kernel density. Excluding them from MVEE’s input should result in better ellipsoidal approximation of the “main” part of the protein’s body. The question remains however how to differentiate between outliers and non-outliers. Because there is no universal cutoff value (each protein has its own density profile), we introduce here threshold $t \geq 0$, derived from a user-controlled parameter $m \geq 0$. Effective atoms for which kernel density falls below t are considered outliers (Figure 1d–f). We call the rest “guides” (guides for MVEE).

Value of t is equal to m -th median of the density profile. Its calculation is similar to how quartiles are obtained, except that we are only moving towards the bottom of the scale. First median ($m = 1$) is calculated for the complete profile ($t = 0.53$ for 1L2P). It is the standard median. Second median ($m = 2$) is the median of all values of the profile not larger than the first median ($t = 0.34$ for 1L2P). Third median ($m = 3$) is the median of all values of the profile not larger than the second median ($t = 0.28$ for 1L2P) and so on. In other words, with increasing values of m , the resulting values of threshold t decrease, causing fewer effective atoms to join the outlier class.

Setting m to 0 implies that all effective atoms are the guides, essentially allowing the KDE phase to be skipped (the corresponding value of t is then 0). It is meant for users who know their molecule and wish to enclose it in MVEE verbatim, for example after manually removing specific parts of it (i.e., marking them to be ignored). On the other hand, $m = 3$ appears to be the first value which highlights most of the outlying regions of the protein but also does not trim too much of its surface elsewhere. We believe it to be the universal default of EP algorithm, particularly useful when working with bulks of structures which did not undergo manual inspection. Its effects on 4B0H, 1L2P and 1B79 are shown in Figure 1, while an additional benchmark for 4B0H utilizing the same methodology and values of m between 0 and 5 is presented in Figure 2.

Figure 2. Influence of value of parameter m on classification of effective atoms of monomer of dUTPase YncF (4B0H) as guides (spheres colored in accordance with their kernel density values) and outliers (green spheres, with kernel density below m -based threshold t): $m = 0/t = 0.00/130$ guides/0 outliers (a); $m = 1/t = 0.55/65$ guides/65 outliers (b); $m = 2/t = 0.40/98$ guides/32 outliers (c), $m = 3/t = 0.27/114$ guides/16 outliers (d); $m = 4/t = 0.19/122$ guides/8 outliers (e) and $m = 5/t = 0.09/126$ guides/4 outliers (f). See Figure 1d for the associated density profile. Values of t and names of subfigures where they are applied are marked on the left of the color bar.

It should be noted here that KDE works on top of smoothing already caused by use of effective atoms. Calculating it for effective atoms again improves the algorithm's running time without sacrificing its precision. By observing patterns in outliers—such as presence along longer sequence segments—one can infer about their significance. Scattered singletons have smaller impact on the output and tend to be contained in the final ellipsoid, passing them along the guides for the next processing step. This also means that higher number of outliers does not necessarily translate into much different MVEE. It is therefore recommended to run the complete EP algorithm twice for the same unknown structure using $m = 0$ and $m = 3$ or 4 and compare the results (see Section 3.4). If there are no drastic changes, such as loss of the globular status, output with $m > 0$ can be accepted. Otherwise user's participation is advised.

3.3. Bounding Ellipsoid and Space Voxelization

In this subsection we are again using monomer of 4B0H as the visual example. Steps of its processing are shown in Figure 3. Presence of outlying C-terminal segment between K118 and G131 (Figure 3a) makes this structure a good subject for demonstrating features of EP algorithm. Other example proteins appear in the next subsection.

Figure 3. Four steps of application of EP algorithm to monomer of dUTPase YncF (4B0H). Step 1 (a): preparation of the structure with atoms shown as van der Waals spheres (blue—towards N-terminus, red—towards C-terminus). Step 2 (b): selection of effective atoms, separation of effective atoms into guide (blue) and outlier (green) subsets using $m = 3$, bounding in a minimum volume ellipsoid. Step 3 (c): alignment with axes of the coordinate system, generation of the grid. Step 4 (d): voxelization of the protein. Colors of voxels on (d) represent their matching values of Equation (1): blue—towards ellipsoid's center, red—towards ellipsoid's surface. Green spheres on (c,d) are atoms located outside the MVEE.

Once effective atoms of the input protein are designated and their kernel density is obtained via KDE, it is time to choose their guide subset and pass it to MVEE. Guide choice is explained in-depth in the previous subsection (to recollect: $m = 3$ is our default). In any case, effective atoms should be now enclosed in their minimum volume ellipsoid, calculated using error margin $\varepsilon = 0.01$ (Figure 3b).

Next step involves shift of the ellipsoid (and the whole protein along with it) to the origin and their rotation in alignment with axes of coordinate system. Translation and rotation operators are obtained from the output of MVEE (see Section 2.2). We also note here lengths of three semi-axes of the ellipsoid (labeled a , b and c), round them to nearest multiple of 1 Å (read: to nearest integer) and increase them by 1 Å. They are needed for the generation of the grid.

Grid is the set of all cubes (voxels) distributed on a uniform lattice with 1 Å steps and located inside the now-axis-aligned ellipsoid. It splits space between a finite number of bins. Because ellipsoid is centered at the origin, we can generate the grid by simply picking all points with integer coordinates $[x, y, z]$ for which value of the ellipsoid's implicit surface equation (e) is below 1 (Figure 3c):

$$e = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \quad (1)$$

$e = 0$ means that $[x, y, z]$ is located at ellipsoid's center, $e = 1$ that it belongs to ellipsoid's surface and $e > 1$ that it lies outside the ellipsoid. We find 1 Å to be a good default grid

resolution, in balance between precision and speed. 2 Å can be used for speed gain with larger structures in exchange for small precision loss (see Section 4).

The penultimate step of EP algorithm involves voxelization of protein's atoms: turning spheres into cubes. It is needed to measure how much of the grid is occupied by the molecule, which is the basis for our main globularity metrics. Voxelization is performed by finding all grid members which have their centers within van der Waals radii of atom centers. For this purpose we use *k*-d tree to do nearest neighbor lookups with NACCESS [70] vdW definitions, which we obtained from dr-sasa sources [71,72]. FreeSASA project provides them too [73,74] among a few alternative configurations. We also employ defaults for residues and atoms not present in that dataset. They follow NACCESS nucleic acid vdW radii and are listed in Table 2. Atoms missing from this table too are assumed to have their vdW radius equal to $\sqrt{3}$. This fallback value allows atoms located at integer coordinates to capture all 27 voxels around it (a complete $3 \times 3 \times 3$ Å cube).

Table 2. Default van der Waals radii for protein voxelization procedure. Hydrogen is provided only for reference. Asterisk (*) denotes default radius for atoms not specified in this table.

Element	Radius	Element	Radius	Element	Radius	Element	Radius
H	1.0	O	1.4	N	1.6	C	1.8
S	1.9	Se	1.9	P	1.9	*	$\sqrt{3}$

The extracted set of grid members resulting from the above subroutine is termed here as the protein's voxels. They are shown for 4B0H in Figure 3d. The protein is now ready for measurement of globularity of its structure.

3.4. Measurements and Comparison

In this section we apply EP algorithm to 1IS9 and other five example proteins (mostly complexes): Ribosomal protein L9 (PDB code: 1DIV, dimer), HLA-DR Invariant Chain (PDB code: 1IIE, trimer), Hypothetical Protein Atu1913 (PDB code: 2B1Y, dimer), Alpha-synuclein (PDB code: 2KKW, monomer) and Uncharacterized Protein (PDB code: 3BPD, heptamer). They present various configurations of input and output. Outcome of their MVEE bounding is presented in Figure 4, while values of globularity metrics for them and other structures from Table 1 is given in Table 3 (using $m = 3$) and in Table 4 (using $m = 0$).

Table 3. Results of application of EP algorithm to selected segments of proteins from Table 1 using $m = 3$. Underline values promote globular status. Underlined PDB codes denote structures deemed globular: single line—partially ($El_{1.0} \geq 0.3$, $0.3 \leq El_{0.3} < 0.5$), double line—fully ($El_{1.0} \geq 0.3$, $El_{0.3} \geq 0.5$). V is ellipsoid's volume coefficient, calculated as $V = a \times b \times c / 1000$. T stands for highest triangle inequality ratio between a , b and c , calculated as $T = \max(a/(b+c), b/(a+c), c/(a+b))$.

PDB Code	Segment (Chains)	Effective Atoms		Ellipsoid Semi-Axes					Ellipsoid Index		Ellipsoid Profile	
		All	Guide	a	b	c	V	T	$El_{0.3}$	$El_{1.0}$	$ EP _{0.1}$	$ EP _{1.0}$
1B79	A – D	410	359	16	41	58	38.05	1.02	0.37	0.28	<u>0.55</u>	<u>0.33</u>
1DIV	A + B	298	261	20	45	51	45.90	<u>0.79</u>	0.21	0.17	0.41	0.20
<u>1IIE</u>	A + B + C	225	197	22	23	23	11.64	<u>0.51</u>	<u>0.68</u>	<u>0.52</u>	<u>0.51</u>	<u>0.60</u>
<u>1IS9</u>	A	358	313	22	24	28	14.78	<u>0.61</u>	<u>0.61</u>	<u>0.58</u>	<u>0.56</u>	<u>0.61</u>
<u>1L2P</u>	A	61	53	6	6	57	2.05	4.75	<u>0.76</u>	<u>0.61</u>	<u>0.88</u>	<u>0.70</u>
<u>2B1Y</u>	A + B	202	177	17	20	42	14.28	1.14	0.48	<u>0.38</u>	0.22	<u>0.44</u>
2KKW	A	140	123	22	36	61	48.31	1.05	0.04	0.06	0.01	0.05
<u>3BPD</u>	A – G	638	558	21	40	41	34.44	<u>0.67</u>	0.24	<u>0.38</u>	0.00	<u>0.31</u>
<u>4B0H</u>	A	130	114	16	22	27	9.50	<u>0.71</u>	<u>0.58</u>	<u>0.40</u>	<u>0.50</u>	<u>0.50</u>

Figure 4. Results of voxelization of six example proteins using $m = 3$: Endoglucanase A (1IS9, monomer) (a), Hypothetical Protein Atu1913 (2B1Y, dimer) (b), HLA-DR Invariant Chain (1IIE, trimer) (c), ribosomal protein L9 (1DIV, dimer) (d), Uncharacterized Protein (3BPD, heptamer) (e), Alpha-synuclein (2KKW, monomer) (f). Colors of voxels represent values of Equation (1): blue—towards ellipsoid’s center, red—towards ellipsoid’s surface. Their range is the same as the range of colors in Figure 3d. Green spheres are atoms of the protein located outside the MVEE.

Table 4. Results of application of EP algorithm to selected segments of proteins from Table 1 using $m = 0$. See description of Table 3 for more information how to interpret this table.

PDB Code	Segment (Chains)	Effective Atoms		Ellipsoid Semi-Axes					Ellipsoid Index		Ellipsoid Profile	
		All	Guide	a	b	c	V	T	$EI_{0.3}$	$EI_{1.0}$	$ EP _{0.1}$	$ EP _{1.0}$
1B79	A – D	410	410	23	41	58	54.69	0.91	0.32	0.23	0.49	0.29
1DIV	A + B	298	298	24	51	52	63.65	0.69	0.23	0.15	0.46	0.20
1IIE	A + B + C	225	225	23	45	45	46.58	0.66	0.38	0.18	0.42	0.29
1IS9	A	358	358	24	26	30	18.72	0.60	<u>0.61</u>	<u>0.53</u>	<u>0.60</u>	<u>0.59</u>
1L2P	A	61	61	6	7	61	2.56	4.69	<u>0.74</u>	<u>0.56</u>	<u>0.88</u>	<u>0.66</u>
2B1Y	A + B	202	202	20	26	51	26.52	1.11	0.30	0.21	0.19	0.26
2KKW	A	140	140	23	43	62	61.32	0.94	0.03	0.05	0.00	0.04
3BPD	A – G	638	638	30	40	41	49.20	0.59	0.20	<u>0.31</u>	0.00	0.25
4B0H	A	130	130	20	24	35	16.80	0.80	0.36	0.24	<u>0.56</u>	<u>0.31</u>

Voxels representing space occupied by atoms of the input protein give an estimate of its volume in $\text{\AA}^3$. Direct comparison (i.e., division) of their number with total number of grid members that fit inside the corresponding MVEE does not fully answer how well the molecule is filling this shape. Due to how MVEE algorithm works, there are locations within almost every ellipsoid which cannot be occupied by atoms, even if the protein fits very well inside it. Voxel counting does not distinguish between this kind of “natural” empty space and cavities in the molecular surface, both on the outside and on the inside. Size of the ellipsoid alone (volume or triangle inequality of its semi-axes) also does not carry enough information to classify it. For instance, 2B1Y (Figure 4b) is considered partially globular by EP, despite the longest semi-axis of this complex being longer (although barely) than the sum of lengths of other semi-axes. On the other hand, 2KKW (Figure 4f) has similar highest triangle inequality ratio (1.14 vs. 1.05), yet it occupies the bottom zone of EP-measured globularity spectrum. They could be differentiated via molecular surface area comparison, but since we are not using it here, we address this issue by giving every point following weight w :

$$w = \max\left(0; 1 - \frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2}\right) \quad (2)$$

Meaning of the symbols is retained from Equation (1). Equation (2) causes center of the ellipsoid to have highest importance ($w = 1$) that decreases with distance down to 0 at its surface and beyond it. In other words, everything outside the MVEE is ignored. This is balanced by the fact that there are fewer atoms near the center than away from it, preventing it from dominating the rest. As a result, this approach allows us to calculate the value of ellipsoid index (EI), our measure of globularity.

Ellipsoid index (EI) is a function of voxels, grid, semi-axes of MVEE and a user-controlled parameter $i = 0 \dots 1$. It controls which protein's voxels and grid members located between the origin and a boundary nominated by this parameter are “indexed” (captured and passed through Equation (2)). Values of Equation (1) calculated for their centers must simply not surpass i . This means that $i = 0$ selects only the point at $[0, 0, 0]$, while $i = 1$ selects the whole ellipsoid. Note that following the previous subsection, there are no grid members (or protein's voxels thereof) on the surface of the ellipsoid and beyond it. Division of sum of weighted volumes of “indexed” protein's voxels by sum of weighted volumes of “indexed” grid members is the value of ellipsoid index at i , written as EI_i . Like Equations (1) and (2), it can attain values between 0 (no protein's voxels captured) to 1 (all protein's voxels captured which are also equal to all grid members captured). $EI_i = 1$ is

therefore possible only when the molecule perfectly fills portion of the ellipsoid that is specified by i . In real structures it may happen only for $EI_{0.0}$.

Calculating EI_i for all i between 0 and 1 yields the ellipsoid profile (EP) of the input protein. We do it in discrete steps of 0.01 which provide an adequate resolution. Profiles of the example proteins are shown in Figure 5 ($m = 3$) and Figure 6 ($m = 0$). Because ellipsoid index is calculated for all points between the origin and the i -based boundary, further values of the profile (closer to the surface) are influenced by previous values (closer to the center). This turns it into a “fingerprint” of the distribution of molecule’s atoms in respect to their enclosing ellipsoid. Encountering filled regions causes the profile to raise, while void lowers it. Owing to the weights (Equation (2)), empty space outside the protein has lower impact on the shape of the distribution, but internal cavities (and structural outliers possibly affecting the MVEE) are still reflected in it. This is how the motivation behind our algorithm is realized. If needed, one may also employ different weight functions.

Figure 5. Ellipsoid profiles (a) and corresponding ellipsoid profile maps (b) of proteins from Table 3 using $m = 3$.

Four colored zones can be seen underneath the profiles in Figures 5a and 6a: red, orange, green and white/transparent. These zones are visual cues for classification and comparison of proteins by the means of EP algorithm. They are coupled with two special values of parameter i : 1.0 and 0.3. $EI_{1.0}$ is the “main” index, reflecting everything that happened along the profile. Our experiments suggest that for a protein to be able to pass as globular, its $EI_{1.0}$ should be not lower than 0.3. Status of $EI_{1.0} < 0.3$ is represented by the red zone. However, $EI_{1.0}$ only accounts for the overall fitting to the ellipsoid. Figure 4e displays a rare counterexample—3BPD. This toroidal heptamer has $EI_{1.0} \geq 0.3$ but at the same time its center is completely devoid of atoms, resulting in an anomalous EP, initially going through the red zone, but eventually raising above it. Such situation can be detected with $EI_{0.3}$ which characterizes center-to-middle globularity of the structure. If it is not below 0.5 (in addition to $EI_{1.0} \geq 0.3$), then the complete input can be assigned a fully globular status. Partial globular status happens when $EI_{1.0} \geq 0.3$ and $0.3 \leq EI_{0.3} < 0.5$. Profiles of highly globular proteins—such as 1IS9 (Figure 4a)—are expected to only intersect the green and white/transparent zones, while those with exposed midsection—like 2B1Y—to also go through the orange zone. Due to dependence of $EI_{1.0}$ on previous indexes, situations where $EI_{1.0} < 0.3$ but $EI_{0.3} \geq 0.5$ are expected to be very rare.

Figure 6. Ellipsoid profiles (a) and corresponding ellipsoid profile maps (b) of proteins from Table 4 using $m = 0$.

It should be noted here that values of i (0.3, 1.0) and their associated thresholds (0.5, 0.3) presented in the previous paragraph were chosen to distinguish structures which under visual inspection appear globular enough from those which do not appear to be globular, i.e., 1IS9 vs. 2KKW. Nonetheless, these numbers are not constant, so depending on the context, one may want to lower or raise the thresholds or scrutinize different indexes, such as $EI_{0.1}$ which reflects the status of the center of the ellipsoid.

Ellipsoid profiles of all proteins exist in the same unit square. By using $EI_{0.3}$ and $EI_{1.0}$ as coordinates, one can plot an ellipsoid index map, allowing visual comparison between the various structures. Such maps for proteins from Table 1 are shown in Figures 5b and 6b. Because of the aforementioned dependency of $EI_{1.0}$ on previous values of the profile, high correlation between it and $EI_{0.3}$ can be seen in these figures (with 3BPD as the outlier). Red, orange and green zones also reprise their roles there.

The initial white/transparent zone in Figures 5a and 6a (between $i = 0.0$ and $i = 0.1$) corresponds to region close to ellipsoid's center which witnesses strongest profile fluctuations. For instance, EP of 1IS9 starts at $EI_{0.0} = 1.0$, while EP of 1IIE starts at $EI_{0.0} = 0.0$. Position of this location depends only on whether the central voxel at $[0, 0, 0]$ is occupied by an atom. If it is not, but there are nearby atoms, the profile should raise quickly, possibly entering the green zone, like it does for 1IIE. It does not happen in 1DIV where conformation of the complex imposes a spacious ellipsoid (Figure 4d), causing the profile to finish in the red zone. Because of this, we do not consider the 0.0–0.1 region as decisive in terms of protein's globularity, but as secondary hint regarding the status of its most inner part. It appears globular to us when area under the corresponding part of the profile—calculated using trapezoidal rule—is not below half of chart area of that region (50% of 0.1). We symbolize this ratio with $|EP|_{0.0-0.1}$, or in short: $|EP|_{0.1}$. Once again, this threshold can be adjusted to fit the specifics of the experiment.

Chart area ratio under ellipsoid profile between $i = 0.1$ and $i = 1.0$ (expressed as percentage of 0.9) is also useful. While it cannot be relied upon for detection of profile's intersection with red and orange zones, $|EP|_{0.1-1.0}$ ($|EP|_{1.0}$ in short) can simplify profile comparison as it binds the indexes under a single value. For example, it marks high discrepancy between the results using $m = 3$ and $m = 0$ for 1IIE complex which are caused by its disordered C-termini (Figure 4c). Conversely, the same kind of difference observed for 1IS9 and 1DIV is negligible, confirming that there are no significant structural outliers

in those proteins. They also demonstrate that such situation may happen in globular and non-globular structures alike. Note however that non-globular and partially globular status applies respectively to complex and chains of 1DIV, but not to its domains.

Last metrics employed by EP algorithm are actually available before the indexes. These are lengths of semi-axes of MVEE (a, b, c) and their derived coefficients: V and T . Together they highlight structures which are deemed globular ($EI_{1.0} \geq 0.3$ and $EI_{0.3} \geq 0.5$) but are intrinsically non-globular. 1L2P—a long helix—is a very good illustration. The three semi-axes of its ellipsoid are: $a = 6, b = 6, c = 57$. Its low volume coefficient (calculated as $V = a \times b \times c / 1000 = 6 \times 6 \times 57 / 1000 = 2.05$) and strong violation of triangle inequality by longest semi-axis (calculated as $T = c / (a + b) = 57 / (6 + 6) = 4.75$) confirm a highly elongated and thin but straight structure. 1IIE with $T = 0.51$ ($m = 3$) is almost spherical. In contrast, high volume ($V = 45.9$) but low $EI_{1.0}$ in 1DIV (0.17) is a signal that its complex possesses a spacious, non-globular conformation (Figure 4d). Same conclusion is reached for 2KKW (Figure 4f), with its $|EP|_{0.1} = 0.01$ additionally confirming void near ellipsoid's center.

3.5. Domain Superfamily Bulk Analysis

In order to demonstrate the capabilities of EP algorithm on a larger dataset, we applied it to domains representing 2124 SCOP 2.08 superfamilies. Their identifiers were obtained from ASTRAL compendium with addition of members of class h (see Section 2.1). These superfamilies represent 1257 folds. 1110 of those folds have only one representative. The rest have close to 7 on average (27 have above 9), with highest number, 62, given to ferredoxin-like (d.58). We believe however such redundancy is permitted here, because even though superfamilies share a common fold, their tertiary structure may be different enough to yield different values of EP measures. They are also located at more diverse (lower) level in SCOP hierarchy, giving a better overview of the whole database.

We ran EP algorithm twice for each domain, using $m = 3$ and $m = 0$, with $T \geq 2$ as the elongation condition (to detect structures similar to 1L2P). The results of this calculation are available in Supplementary Materials, while their summary is given in Tables 5 and 6 and visualized in Figure 7, split between the eight SCOP classes (a through h). We decided to limit the main graphical output to “main” m value (3), because as expected, $m = 0$ causes a nearly universal decrease in $EI_{0.3}$ and $EI_{1.0}$ but without immediately noticeable changes on the profile map. We present however differences between $|EP|_{1.0}$ for $m = 3$ and $m = 0$ in Figure 7i as a sorted line plot, which is suitable for showing their magnitude.

Table 5. Results of application of EP algorithm to SCOP 2.08 domain superfamily representatives using $m = 3$. CL—SCOP class, SF—number of superfamilies, CF—number of common folds, Guides—average number of guide effective atoms, Volume—average ellipsoid volume coefficient (V), $EI_{0.3}$ and $EI_{1.0}$ —average ellipsoid indexes, CC— $EI_{0.3}$ vs. $EI_{1.0}$ correlation coefficient, R—number of “red” domains (non-globular), O—number of “orange” domains (partially globular), G—number of “green” domains (globular), $|EP|_{0.1+}$ —number of domains where $|EP|_{0.1} \geq 0.5$ (globular center of the structure), $EI_{1.0+}$ —number of domains where $EI_{1.0} > EI_{0.3}$, T+—number of domains with T coefficient larger or equal 2 (elongated structure). Right columns under Guides, Volume, $EI_{0.3}$ and $EI_{1.0}$ list standard deviations for mean values to their left.

CL	SF	CF	Guides		Volume		$EI_{0.3}$		$EI_{1.0}$		CC	R	O	G	$ EP _{0.1+}$	$EI_{1.0+}$	T+
<i>a</i>	519	290	117	79	8.2	7.1	0.56	0.12	0.46	0.10	0.90	47	62	410	364	29	14
<i>b</i>	374	179	142	88	9.4	8.6	0.58	0.09	0.48	0.09	0.86	21	21	332	286	9	2
<i>c</i>	246	147	204	92	12.5	7.3	0.58	0.07	0.49	0.07	0.84	4	17	225	212	2	0
<i>d</i>	577	395	123	68	8.2	7.9	0.58	0.10	0.48	0.09	0.89	36	40	501	471	8	3
<i>e</i>	73	73	315	178	27.1	26.2	0.50	0.11	0.40	0.09	0.92	13	15	45	49	2	0
<i>f</i>	130	69	154	126	13.4	12.9	0.50	0.18	0.40	0.13	0.94	31	31	68	70	12	34
<i>g</i>	139	98	54	27	3.4	3.0	0.59	0.14	0.49	0.12	0.93	11	13	115	113	5	1
<i>h</i>	66	6	92	73	10.6	17.1	0.52	0.25	0.43	0.18	0.98	15	12	39	38	11	46
all	2124	1257	137	98	9.7	10.4	0.57	0.12	0.47	0.10	0.91	178	211	1735	1603	78	100

Table 6. Results of application of EP algorithm to SCOP 2.08 domain superfamily representatives using $m = 0$. Average number of guide effective atoms is for $m = 0$ equal to the average number of all residues in the domains. See description of Table 5 for more information how to interpret this table.

CL	SF	CF	Guides		Volume		$El_{0.3}$		$El_{1.0}$		CC	R	O	G	$ EP _{0.1+}$	$El_{1.0+}$	T+
a	519	290	134	90	11.3	9.8	0.52	0.13	0.39	0.10	0.90	91	77	351	370	24	8
b	374	179	162	101	13.1	12.1	0.56	0.11	0.41	0.09	0.89	45	26	303	291	5	1
c	246	147	233	105	16.8	10.0	0.56	0.08	0.43	0.07	0.87	13	31	202	200	1	0
d	577	395	141	78	11.2	10.7	0.55	0.11	0.41	0.09	0.91	81	49	447	461	5	1
e	73	73	360	204	35.5	31.4	0.47	0.11	0.35	0.09	0.91	22	16	35	46	1	0
f	130	69	176	144	18.2	17.0	0.44	0.17	0.33	0.12	0.94	52	27	51	64	15	26
g	139	98	61	31	4.6	3.9	0.56	0.14	0.43	0.12	0.94	23	14	102	110	2	0
h	66	6	105	83	14.6	20.9	0.44	0.23	0.35	0.16	0.97	30	7	29	36	13	44
all	2124	1257	157	113	13.1	13.6	0.53	0.13	0.40	0.10	0.92	357	247	1520	1578	66	80

Figure 7. Ellipsoid profile maps of domains representing SCOP 2.08 superfamilies in all alpha (a), all beta (b), alpha/beta (c), alpha+beta (d), multi-domain (e), membrane and cell surface (f), small (g) and coiled coil (h) classes. Color of the markers matches color of the zone they are occupying (red, orange, green). Size of the markers denotes total number of residues, clipped to 50–200 range. Shape of the markers signals value of T coefficient: circle—less than 2 (ellipsoidal shape), square—2 or more (elongated shape). Blue diamonds mark position of average values of $El_{0.3}$ and $El_{1.0}$. Data ranges of axes on (a–h) are limited to 0.9 and 0.8 respectively for visibility. There are only two domains which lie outside them: d5nw3a_ (0.91×0.82 , class g) and d1ca4a2 (0.91×0.80 , class h). Chart on (i) compares $|EP|_{1.0}$ values between $m = 3$ and $m = 0$.

On average, $m = 0$ lowered $|EP|_{1.0}$ by 0.05 ($\sigma = 0.04$, $\max = 0.33$, $\min = -0.03$), by more than 0.1 in 184 domains and by more than 0.2 in 28 domains (least in classes *b* and *c*, none in *e*). Some of them are small, bent helices, like d1eq7a_. Only in 39 domains $m = 0$ caused a tiny increase in $|EP|_{1.0}$ (again, none in class *e*). $m = 3$ vs. $m = 0$ $|EP|_{1.0}$ correlation coefficient is 0.94. In total, $m = 0$ caused 215 domains to lose globular status.

Further analysis focuses only on results for $m = 3$. From Figure 7 and Tables 5 and 6 one can see that there is a high correlation between $EI_{0.3}$ and $EI_{1.0}$ (0.85 and above) in all classes. Most domains are also considered globular by EP. Only 8% of them was assigned to red zone and 10% to orange zone. This is accordant with our expectations and a cursory look at 3D structure of those domains reveals correct assessment. One such example is d2es4d1, member of SCOP non-globular all-alpha subunits of globular proteins superfamily (a.137 fold), which wraps around other parts of the molecule. We can confirm with our algorithm that SCOP domains are predominantly globular.

It appears that 0.7 and 0.6 are the practical top limits for $EI_{0.3}$ and $EI_{1.0}$ respectively. Only 76 domains reach $EI_{1.0} > 0.6$, 61 reach $EI_{0.3} > 0.7$ and 43 reach both. Most of them are found in *f*, *g* and *h* classes. Small proteins (class *g*) are generally given globular rating (Figure 7g). On the other hand, the two least globular domains are also in this class: d6y3ba_ and d2ffta1, which SCOP unsurprisingly categorize to non-structural (g.96) and intrinsically disordered (g.88) folds. Returning to the opposite side, the most globular structures in *f* and *h* classes are small, simple helices (70% of class *h* has $T \geq 2$). This is why they show as squares in Figure 7f,h. d1dpjb_ (a.137 once again), located in similar region in Figure 7a, is also a straight helix. Interestingly, this protein's complex has similar shape to d5ae0a_ (d.61, LigT-like), the "best" from class *d*. Highest globularity in beta-only proteins is found for d6s2ma1 (b.60, lipocalin, hydrophobic ligand transport) and d1ezga_ (b.80, cysteine-rich antifreeze protein). Class *c* (alpha/beta) has least number of "red" domains, only four. Its biggest outlier, d1vq8l1 represents SCOP ribosomal proteins L15p and L18e fold and exhibits the same shape feature as d2es4d1.

When size of domains, measured via number of residues and ellipsoid's volume, is analyzed, class *e* stands out, doubling the average length and volume of other classes. Its corresponding chart (Figure 7e) is also compact and has no outliers. Other type of outliers, where $EI_{1.0}$ is low but higher than $EI_{0.3}$, mostly appear in class *a*, but domains with highest difference between these indexes also exist in classes *b* and *d*. These are: d2ag4a2 (b.95, Ganglioside M2 activator), d1pprm1 (a.131, Peridinin-chlorophyll protein) and d5njcb_ (d.283, Putative modulator of DNA gyrase). Like 3BPD complex, they assume a bowl-like shape, fitting well to MVEE but with hollow middle.

Based on the above analysis, we can confirm that with EP algorithm one can estimate and study structural features of input proteins in respect to their impact on overall globularity and which are in accordance with human observer's expectations towards them.

4. Discussion and Conclusions

In this paper we introduced the Ellipsoid Profile (EP) algorithm for simple measurement of general globularity in protein structures. It encloses its input molecule in minimum volume enclosing ellipsoid (MVEE) and checks how well atoms of the query fill this shape. Below is a brief summary of its steps:

1. Load and prepare the input structure (Section 3.1).
2. Designate effective atoms of residues selected for calculation (Section 3.2).
3. Calculate density of effective atoms using kernel density estimation (Section 3.2).
4. Choose guide effective atoms—those with density above threshold t , selected upon the basis of user-controlled parameter $m \geq 0$ (Section 3.2).
5. Fit MVEE to guide effective atoms using parameter $\varepsilon \leq 0.01$ (Section 3.3).
6. Using output of MVEE, shift the resulting ellipsoid and all protein's atoms to origin and rotate them in alignment with axes of coordinate system (Section 3.3).
7. Generate uniform grid inside the axis-aligned ellipsoid (Section 3.3).

8. Voxelize the protein by finding grid neighbors of its atoms located within vdW radii of those atoms (Section 3.3).
9. Calculate globularity metrics (Section 3.4).

EP algorithm provides following synergistic measures of globularity, which can be used for comparison and classification of protein structures based on this property:

- Ellipsoid indexes, including two “standard” ones: $EI_{0.3}$ and $EI_{1.0}$;
- Ellipsoid profile (EP), the distribution of EI_i for $i = 0...1$ in 0.01 intervals;
- Areas under profile: $|EP|_{0.1}$ (for $i = 0.0...0.1$) and $|EP|_{1.0}$ (for $i = 0.1...1.0$);
- Lengths of semi-axes of MVEE and their derivatives: volume coefficient (V) and highest triangle inequality ratio (T).

It is not necessary to focus just on $EI_{0.3}$ and $EI_{1.0}$, but in our opinion and on the basis of presented results, they provide a solid foundation for measurement of globularity. Interpretation of their values is also subject to researcher’s needs. One may increase or decrease their associated thresholds (see below), for example to bring out only the most globular structures. Our default interpretation is as follows:

- Structure with $EI_{1.0} < 0.3$ or $EI_{0.3} < 0.3$ is non-globular.
- Structure with $EI_{1.0} \geq 0.3$ and $0.3 \leq EI_{0.3} < 0.5$ is partially globular.
- Structure with $EI_{1.0} \geq 0.3$ and $EI_{0.3} \geq 0.5$ is globular.
- Structure with $|EP|_{0.1} \geq 0.5$ has globular center (secondary metric).
- Structure with $|EP|_{0.1} < 0.5$ may have a central cavity.
- Structure with $T < 1$ has ellipsoidal shape (spherical MVEE if $T = 0.5$).
- Structure with $T \geq 2$ has elongated shape (probably fibrous protein).

EP algorithm can be implemented using common software packages. This allows it to be directly added to protein processing pipeline in another program. Its running time is also short despite our Python reference implementation (although owing to optimized C libraries), measured in milliseconds even for medium structures on a modest CPU of a laptop computer (i.e., for 3BPD, 7 chains, over 600 residues). The most time-consuming part is voxelization, which takes as input atoms that have many contacts with the grid. It can be accelerated by increasing grid resolution to 2 Å at the expense of output’s precision. For instance, in 3BPD complex, if MVEE radii are retained, this changes $EI_{0.3}$ from 0.244 to 0.235 and $EI_{1.0}$ from 0.376 to 0.375. In opposite direction (with 0.5 Å grid), $EI_{0.3}$ becomes 0.243 and $EI_{1.0}$ becomes 0.377, but KDT search takes four times longer. Higher resolution might be beneficial when working with small proteins, but one must consider its cost.

Various structures analyzed in Section 3 confirm that EP algorithm is capable of smoothing the molecular surface without actually resolving it. Small, usual irregularities are handled by effective atom representation and larger outliers (protruding from denser “bases”) can be found and eliminated via KDE. This produces a tight, more natural from human observer’s perspective ellipsoidal fit to the structure. Number of medians (m) parameter allows fine-tuning of strength of KDE-based outlier detection, although its default (3) seems to work well for all presented cases and SCOP domains. It is also the only critical parameter of our algorithm. Comparison with results for $m = 0$ (no KDE) can gauge whether there are indeed significant outliers in the structure.

It is worth reminding here that EP algorithm measures only the general globularity of its input. Most of SCOP 2.08 superfamily representatives ($\approx 80\%$) are indeed globular in this sense. The rest that is deemed not globular has valid reasons for it, such as being thin wrappers around other chains. The base algorithm is however unable to distinguish—for example—between a globular alpha-only domain and a membrane protein with similar shape. It only checks which of them resemble a globule, although we provide additional means of detecting fibrous structures via T and V metrics.

Only proteins are considered here, but robustness of EP algorithm allows it to be used with any structure type (i.e., a protein-nucleic complex) or even with non-molecular input, like SASA calculators which require only coordinates and sphere radii. We see primary application of our algorithm as a support tool for another program. Beneficiaries include

hydrophobic core analyzers like FOD model (which similarly revolves around the concept of ellipsoid) and domain explorers, especially those looking for domains that span more than one chain. It may also help in prediction of protein–protein contact areas.

Finally, we would like to speculate about practical role of EP algorithm as scoring function in optimization process where one is—for example—searching for globular parts of the input structure. For highest performance, this structure should be first aligned via MVEE, with its grid and voxels selected as usual. Subsequent ellipsoid index calculations for residue subset candidates should call MVEE on the effective atoms but instead capture the initial voxels and grid members using quadric form equation of the ellipsoid (with matrix M and vector c). This will yield slightly different results but will also be much faster as it circumvents the most time-consuming nearest neighbor search.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/cryst11121539/s1>. Results of application of Ellipsoid Profile algorithm to 2124 SCOP 2.08 domain superfamily representatives using $m = 3$ and $m = 0$. Files are in TSV format with UNIX newline characters. Columns: PDB_ID: PDB code of the structure, SCOP_ID: SCOP entry of the domain, SCOP_CODE: SCOP classification of the domain, SELECTION: chain ID and residue range of the domain, M: value of m parameter, ALL: total number of effective atoms, GUIDE: number of guide effective atoms, A: length of the first semi-axis of MVEE, B: length of the second semi-axis of MVEE, C: length of the third semi-axis of MVEE, V: value of volume coefficient, T: value of highest triangle inequality coefficient, EI(0.3): value of $EI_{0.3}$, EI(1.0): value of $EI_{1.0}$, $|EP|(0.1)$: value of $|EP|_{0.1}$, $|EP|(1.0)$: value of $|EP|_{1.0}$, SEQUENCE: sequence of the domain (only available residues; lowercase letters mark outliers; names of non-standard residues with no parent data in MODRES records are enclosed in parenthesis).

Funding: This research was funded by Jagiellonian University Medical College grant number N41/DBS/000719.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Online calculation of Ellipsoid Profile algorithm and related data is available at <http://fod.cm-uj.krakow.pl> web server.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Brenner, S.E.; Chothia, C.; Hubbard, T.J.; Murzin, A.G. Understanding protein structure: Using scop for fold interpretation. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 1996; Volume 266, pp. 635–643. [CrossRef]
2. Hou, J.; Sims, G.E.; Zhang, C.; Kim, S.-H. A global representation of the protein fold space. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 2386–2390. [CrossRef]
3. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
4. Gish, W.; States, D.J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **1993**, *3*, 266–272. [CrossRef]
5. Available online: <https://blast.ncbi.nlm.nih.gov> (accessed on 7 November 2021).
6. The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2018**, *47*, D506–D515. [CrossRef]
7. Available online: <https://www.uniprot.org> (accessed on 7 November 2021).
8. Berman, H.M.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
9. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichtlow, G.V.; Christie, C.H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J.M.; et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2020**, *49*, D437–D451. [CrossRef] [PubMed]
10. Available online: <https://www.rcsb.org> (accessed on 7 November 2021).
11. Hou, J.; Jun, S.-R.; Zhang, C.; Kim, S.-H. From The Cover: Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 3651–3656. [CrossRef] [PubMed]

12. Banach, M.; Konieczny, L.; Roterman, I. The fuzzy oil drop model, based on hydrophobicity density distribution, generalizes the influence of water environment on protein structure and function. *J. Theor. Biol.* **2014**, *359*, 6–17. [CrossRef] [PubMed]
13. Dułak, D.; Gadzała, M.; Stapor, K.; Fabian, P.; Konieczny, L.; Roterman, I. Folding with active participation of water. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 13–26. [CrossRef]
14. Konieczny, L.; Roterman, I. Information encoded in protein structure. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 27–39. [CrossRef]
15. Banach, M.; Konieczny, L.; Roterman, I. Composite structures. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 117–133. [CrossRef]
16. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540. [CrossRef]
17. Chandonia, J.-M.; Fox, N.K.; Brenner, S.E. SCOPe: Classification of large macromolecular structures in the structural classification of proteins—Extended database. *Nucleic Acids Res.* **2018**, *47*, D475–D481. [CrossRef]
18. Fox, N.K.; Brenner, S.E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2013**, *42*, D304–D309. [CrossRef]
19. Available online: <https://scop.berkeley.edu> (accessed on 7 November 2021).
20. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V.P.; Ashford, P.; Scholes, H.M.; Pang, C.S.M.; Woodridge, L.; Rauer, C.; Sen, N.; et al. CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **2020**, *49*, D266–D273. [CrossRef] [PubMed]
21. Lewis, T.E.; Sillitoe, I.; Dawson, N.; Lam, S.D.; Clarke, T.; Lee, D.; Orengo, C.; Lees, J. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **2017**, *46*, D435–D439. [CrossRef] [PubMed]
22. Available online: <https://www.cathdb.info> (accessed on 7 November 2021).
23. Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A.G. SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Res.* **2013**, *42*, D310–D314. [CrossRef] [PubMed]
24. Kalinowska, B.; Banach, M.; Wiśniowski, Z.; Konieczny, L.; Roterman, I. Is the hydrophobic core a universal structural element in proteins? *J. Mol. Model.* **2017**, *23*, 205. [CrossRef]
25. Konieczny, L.; Roterman, I. Globular or ribbon-like micelle. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 41–54. [CrossRef]
26. Kalinowska, B.; Banach, M.; Konieczny, L.; Roterman, I. Application of Divergence Entropy to Characterize the Structure of the Hydrophobic Core in DNA Interacting Proteins. *Entropy* **2015**, *17*, 1477–1507. [CrossRef]
27. Roterman, I.; Banach, M.; Konieczny, L. Application of the Fuzzy Oil Drop Model Describes Amyloid as a Ribbonlike Micelle. *Entropy* **2017**, *19*, 167. [CrossRef]
28. Banach, M.; Konieczny, L.; Roterman, I. The Amyloid as a Ribbon-Like Micelle in Contrast to Spherical Micelles Represented by Globular Proteins. *Molecules* **2019**, *24*, 4395. [CrossRef]
29. Konieczny, L.; Roterman, I. Description of the fuzzy oil drop model. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 1–11. [CrossRef]
30. Banach, M.; Konieczny, L.; Roterman, I. The active site in a single-chain enzyme. In *From Globular Proteins to Amyloids*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 71–78. [CrossRef]
31. Banach, M.; Chomilier, J.; Roterman, I. Contribution to the Understanding of Protein–Protein Interface and Ligand Binding Site Based on Hydrophobicity Distribution—Application to Ferredoxin I and II Cases. *Appl. Sci.* **2021**, *11*, 8514. [CrossRef]
32. Dygut, J.; Kalinowska, B.; Banach, M.; Piwowar, M.; Konieczny, L.; Roterman, I. Structural Interface Forms and Their Involvement in Stabilization of Multidomain Proteins or Protein Complexes. *Int. J. Mol. Sci.* **2016**, *17*, 1741. [CrossRef]
33. Liu, Q.; Wang, P.-S.; Zhu, C.; Gaines, B.B.; Zhu, T.; Bi, J.; Song, M. OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein–ligand affinity prediction. *J. Mol. Graph. Model.* **2021**, *105*, 107865. [CrossRef] [PubMed]
34. Mylonas, S.K.; Axenopoulos, A.; Daras, P. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **2021**, *37*, 1681–1690. [CrossRef] [PubMed]
35. Prabhakaran, M.; Ponnuswamy, P.K. Shape and surface features of globular proteins. *Macromolecules* **1982**, *15*, 314–320. [CrossRef]
36. Han, X.; Sit, A.; Christoffer, C.; Chen, S.; Kihara, D. A global map of the protein shape universe. *PLoS Comput. Biol.* **2019**, *15*, e1006969. [CrossRef] [PubMed]
37. Wu, H.; Zhang, R.; Zhang, W.; Hong, J.; Xiang, Y.; Xu, W. Rapid 3-dimensional shape determination of globular proteins by mobility capillary electrophoresis and native mass spectrometry. *Chem. Sci.* **2020**, *11*, 4758–4765. [CrossRef] [PubMed]
38. Osadchy, M.; Kolodny, R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12301–12306. [CrossRef] [PubMed]
39. Erickson, H.P. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol. Proced. Online* **2009**, *11*, 32–51. [CrossRef]
40. Available online: <https://scop.berkeley.edu/astrol/subsets> (accessed on 7 November 2021).
41. Available online: <https://files.rcsb.org/pub/pdb/data/status/obsolete.dat> (accessed on 7 November 2021).
42. Fass, D.; E Bogden, C.; Berger, J.M. Crystal structure of the N-terminal domain of the DnaB hexameric helicase. *Structure* **1999**, *7*, 691–698. [CrossRef]
43. Hoffman, D.; Davies, C.; Gerchman, S.; Kycia, J.; Porter, S.; White, S.; Ramakrishnan, V. Crystal structure of prokaryotic ribosomal protein L9: A bi-lobed RNA-binding protein. *EMBO J.* **1994**, *13*, 205–212. [CrossRef]

44. Jasanoff, A.; Wagner, G.; Wiley, D.C. Structure of a trimeric domain of the MHC class II-associated chaperonin and targeting protein Ii. *EMBO J.* **1998**, *17*, 6812–6818. [CrossRef]
45. Schmidt, A.; Gonzalez, A.; Morris, R.J.; Costabel, M.; Alzari, P.M.; Lamzin, V.S. Advantages of high-resolution phasing: MAD to atomic resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 1433–1441. [CrossRef]
46. Del Rizzo, P.A.; Bi, Y.; Dunn, A.S.D.; Shilton, B.H. The “Second Stalk” of Escherichia coli ATP Synthase: Structure of the Isolated Dimerization Domain. *Biochemistry* **2002**, *41*, 6875–6884. [CrossRef]
47. Nocek, B.; Skarina, T.; Edwards, A.; Savchenko, A.; Joachimiak, A. Crystal Structure of Protein of Unknown Function ATU1913 from *Agrobacterium tumefaciens* str. C58. 2005. Available online: https://www ww p d b . o r g / p d b ? i d = p d b _ 0 0 0 0 2 b 1 y (accessed on 7 November 2021).
48. Rao, J.N.; Jao, C.C.; Hegde, B.G.; Langen, R.; Ulmer, T.S. A Combinatorial NMR and EPR Approach for Evaluating the Structural Ensemble of Partially Folded Proteins. *J. Am. Chem. Soc.* **2010**, *132*, 8657–8668. [CrossRef] [PubMed]
49. Eswaramoorthy, S.; Burley, S.; Sauder, J.; Swaminathan, S. Crystal Structure of an Uncharacterized Protein (O28723_ARCFU) from *Archaeoglobus fulgidus*. 2008. Available online: https://www ww p d b . o r g / p d b ? i d = p d b _ 0 0 0 0 3 b p d (accessed on 7 November 2021).
50. García-Nafria, J.; Timm, J.; Harrison, C.; Turkenburg, J.P.; Wilson, K.S. Tying down the arm in Bacillus dUTPase: Structure and mechanism. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69*, 1367–1380. [CrossRef] [PubMed]
51. Khachiyan, L.G. Rounding of Polytopes in the Real Number Model of Computation. *Math. Oper. Res.* **1996**, *21*, 307–320. [CrossRef]
52. Sun, P.; Freund, R.M. Computation of Minimum-Volume Covering Ellipsoids. *Oper. Res.* **2004**, *52*, 690–706. [CrossRef]
53. Kumar, P.; Yildirim, E.A. Minimum-Volume Enclosing Ellipsoids and Core Sets. *J. Optim. Theory Appl.* **2005**, *126*, 1–21. [CrossRef]
54. Todd, M.J.; Yildirim, E.A. On Khachiyan’s algorithm for the computation of minimum-volume enclosing ellipsoids. *Discret. Appl. Math.* **2007**, *155*, 1731–1744. [CrossRef]
55. Available online: <https://stackoverflow.com/questions/14016898/port-matlab-bounding-ellipsoid-code-to-python> (accessed on 1 October 2021).
56. Available online: <https://www.mathworks.com/matlabcentral/fileexchange/9542-minimum-volume-enclosing-ellipsoid> (accessed on 1 October 2021).
57. Bærentzen, J.A.; Gravesen, J.; Anton, F.; Aanaes, H. Convex Hulls. In *Guide to Computational Geometry Processing*; Springer: London, UK, 2012; pp. 227–240. [CrossRef]
58. Barber, C.B.; Dobkin, D.P.; Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **1996**, *22*, 469–483. [CrossRef]
59. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burocski, E.; Peterson, W.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
60. Bentley, J.L. Multidimensional binary search trees used for associative searching. *Commun. ACM* **1975**, *18*, 509–517. [CrossRef]
61. Maneewongvatana, S.; Mount, D.M. *On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 842–851. [CrossRef]
62. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
63. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
64. Gramacki, A. Nonparametric Density Estimation. In *Nonparametric Kernel Density Estimation and Its Computational Aspects*; Springer: Cham, Switzerland, 2017; pp. 7–24. [CrossRef]
65. Gramacki, A. Kernel Density Estimation. In *Nonparametric Kernel Density Estimation and Its Computational Aspects*; Springer: Cham, Switzerland, 2017; pp. 25–62. [CrossRef]
66. Sullivan, C.; Kaszynski, A. PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *J. Open Source Softw.* **2019**, *4*, 1450. [CrossRef]
67. Available online: <https://vtk.org> (accessed on 7 November 2021).
68. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
69. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
70. Hubbard, S.; Thornton, J. *NACCESS, Computer Program*; Department of Biochemistry Molecular Biology, University College London: London, UK, 1993.
71. Ribeiro, J.; Ríos-Vera, C.; Melo, F.; Schüller, A. Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. *Bioinformatics* **2019**, *35*, 3499–3501. [CrossRef] [PubMed]
72. Available online: https://github.com/nioroso-x3/dr_sasa_n (accessed on 1 October 2021).
73. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research* **2016**, *5*, 189. [CrossRef]
74. Available online: <https://github.com/mittinatten/freesasa> (accessed on 1 October 2021).