Topological Data Analysis of Protein Structures via Persistent Homology

Jingman Wang

1 Introduction

Understanding the three-dimensional structure of proteins is central to elucidating their biological function, molecular mechanisms, and evolutionary relationships. Traditional geometric and statistical methods—such as secondary structure assignment, contact map analysis, and molecular surface computation—are effective at characterizing local motifs and global folding patterns. However, these approaches often fall short in quantifying higher-order connectivity, spatial organization, and multiscale cavities or loops, which may play critical roles in protein function and dynamics.

A key obstacle in protein structure analysis lies in capturing both local and global topological features in a manner robust to noise and deformation. Many conventional descriptors are sensitive to structural perturbations and may miss subtle, yet functionally relevant, topological differences, such as rare, long-lived loops or cavities embedded within otherwise similar folds. Furthermore, a lack of quantitative tools for comparing complex shapes hampers our ability to systematically relate structural differences to biological function.

Persistent homology, a foundational tool in topological data analysis (TDA), provides a rigorous mathematical framework for tracking the emergence and disappearance of topological features, such as connected components (H_0), loops (H_1), and voids (H_2, across multiple spatial scales. This approach generates concise descriptors (barcodes, persistence diagrams) that capture both the abundance and longevity of such features, offering robustness to noise and geometric perturbations. While persistent homology has shown promise in a variety of scientific domains, its systematic application to large-scale protein structure comparison and biological interpretation remains relatively underexplored.

In this study, we address these challenges by conducting a persistent homology-based comparative analysis of three representative proteins: hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP). Our methodology involves (i) extracting atomic point clouds from each protein, (ii) computing persistence barcodes and diagrams to quantify and visualize topological features in H_0, H_1, and H_2, (iii) mapping the spatial origin of the most persistent features, and (iv) using Wasserstein distance to quantify topological similarity between structures. This comprehensive approach enables us to identify both shared and distinct topological fingerprints, relate persistent features to structural regions of interest, and provide new insights into the geometric and biological complexity of protein molecules.

Our results demonstrate that persistent homology not only discriminates global differences in molecular architecture, such as compactness versus openness, but also reveals rare, long-lived loops and cavities potentially linked to protein function. By integrating quantitative topology, spatial visualization, and rigorous similarity measures, our study highlights the unique strengths of TDA as a complementary tool in structural bioinformatics.

2 Related Work

The quantitative analysis of protein structures has a long history, with classical approaches focusing on secondary structure assignment [3], contact map and distance matrix analysis [6], and global geometric descriptors such as surface area or packing density [5]. These methods have significantly advanced our understanding of protein folding, classification, and function. However, most traditional techniques primarily capture local or pairwise geometric relationships, leaving higher-order connectivity and global spatial organization less explored.

Recent years have seen the rise of topological data analysis (TDA) as a powerful framework for extracting multiscale geometric and topological features from complex biological data. In particular, persistent homology has been successfully applied to protein structure analysis, providing descriptors that encode the emergence and persistence of connected components, loops, and cavities across filtration scales [8, 4]. Such

topological fingerprints have enabled new approaches to structure comparison, fold classification, flexibility assessment, and binding site prediction [2, 7].

Despite these advances, several challenges remain. The biological interpretation of persistent features, especially long-lived H_1 (loops) and H_2 (voids), is still an open question, and the integration of topological summaries with spatial mapping and functional annotation is an active area of research [1]. Moreover, the robustness of persistent homology to structural noise, sampling artifacts, and the choice of filtration parameters continues to be investigated.

Our study builds on this literature by applying persistent homology to a comparative analysis of three functionally distinct proteins. We combine barcode and persistence diagram computation with spatial mapping of topological features and quantitative similarity assessment using Wasserstein distance. This approach provides a systematic, interpretable, and biologically relevant framework for protein structure comparison.

3 Methodology

This section outlines the complete analytical workflow employed to extract, process, and topologically characterize three-dimensional protein structures using persistent homology and related techniques. Our methodology is designed to capture both global and local topological features from raw structural data, enabling robust comparison between distinct protein architectures.

3.1 Data Acquisition and Preprocessing

Protein structures were sourced from the RCSB Protein Data Bank (PDB), focusing on three representative proteins: hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP). For each protein, the corresponding PDB file was downloaded, and non-hydrogen atomic coordinates were extracted using the `Bio.PDB` Python library. This resulted in a point cloud $\mathcal{P} = \{x_i \in R^3\}_{i=1}^N$ for each protein, where N is the number of heavy atoms.

To mitigate computational costs in downstream topological computations, random subsampling was applied if the number of atoms exceeded a preset threshold (e.g., 2,000 points). Each point cloud was then centered by subtracting its centroid and, if necessary, scaled for normalization.

3.2 Persistent Homology Pipeline

For each processed point cloud, we computed the Vietoris-Rips filtration, a common simplicial complex used in topological data analysis (TDA). Specifically, for a sequence of increasing scale parameters $\{\epsilon_k\}$, the Vietoris-Rips complex $VR(\mathcal{P}, \epsilon_k)$ contains all k-simplices whose vertices are pairwise within ϵ_k distance. This filtration encodes the evolution of topological features (connected components, loops, voids) as ϵ increases.

Using the `GUDHI` and `ripser` libraries, we computed persistent homology in dimensions H_0, H_1, and H_2 for each protein. The resulting persistence diagrams and barcode plots record the "birth" and "death" scales (b_i, d_i) of topological features, with lifetimes $\ell_i = d_i - b_i$ indicating their prominence. Both diagrams and barcodes were generated for qualitative and quantitative interpretation.

To further summarize topological complexity, we computed Betti curves, which track the number of active k-dimensional features (Betti numbers) as a function of the filtration parameter. Persistence entropy, quantifying the diversity of lifetimes, was also evaluated for each protein and homology dimension.

3.3 Quantitative Comparison and Statistical Analysis

For each protein and homology dimension, we recorded statistics including the number of topological features, maximum and top-k persistence values, and distribution of feature lifetimes. Histograms of H_1 (loop) and H_2 (void) lifetimes were constructed to reveal the prevalence of short-lived versus long-lived features.

To measure the similarity or dissimilarity between proteins, pairwise Wasserstein distances were computed between the H_1 persistence diagrams. These distances provide a rigorous quantification of topological fingerprint differences.

To localize the origin of persistent features, we mapped atoms involved in the formation of the most persistent H_1 (loops) and H_2 (voids) features. Specifically, atoms contributing to the birth of top k persistent features were visualized in the context of the protein's 3D structure, enabling interpretation of their structural or functional relevance.

4 Results

4.1 Atomic Point Cloud Visualization

The three-dimensional atomic point clouds of hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP) are illustrated in Figure 1. Each protein structure demonstrates a unique spatial arrangement of non-hydrogen atoms, reflecting distinct molecular architectures and compactness.

4HHB (hemoglobin) appears as a compact, globular point cloud, consistent with its well-known quaternary structure. 1CRK (creatine kinase) displays a more anisotropic, extended, and cross-shaped distribution, indicating a less compact and potentially more open architecture. 1ATP (protein kinase A) presents an intermediate level of compactness, with atoms moderately dispersed around the center.

These visualizations highlight the intrinsic differences in global geometry among the three proteins, providing a structural basis for subsequent topological analyses.

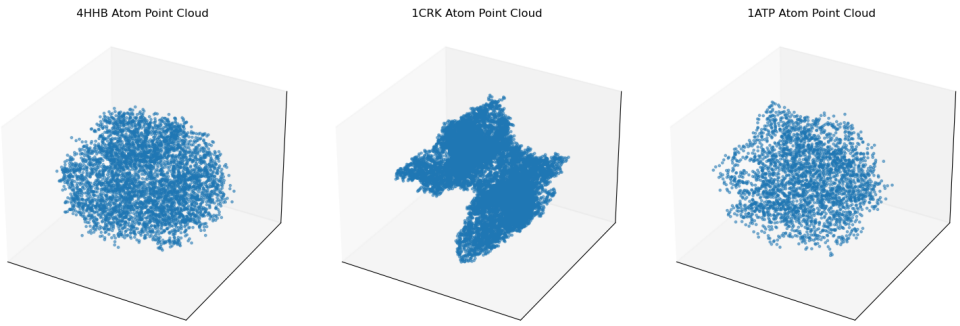


Figure 1: Three-dimensional atomic point clouds for hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP).

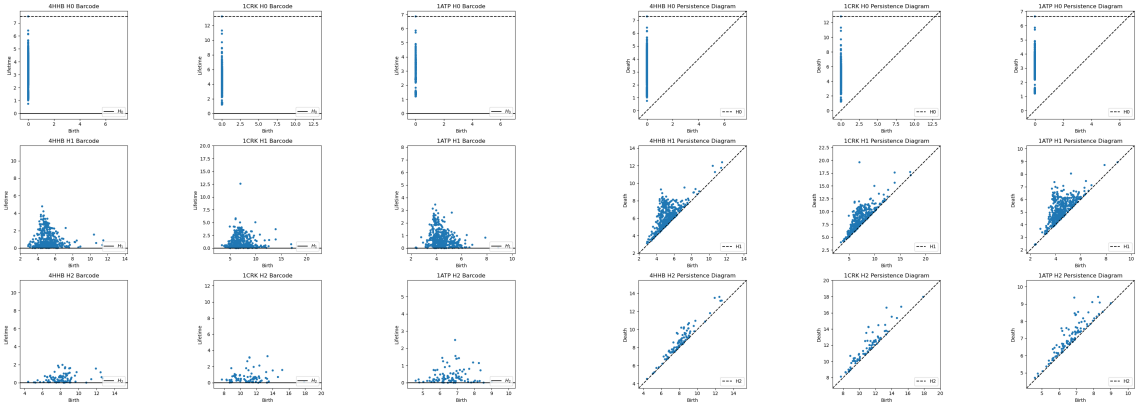
4.2 Topological Summary

To systematically characterize the topological structures present in each protein point cloud, we computed persistence barcodes and persistence diagrams in dimensions H_0 (connected components), H_1 (loops), and H_2 (voids). Figure 2 displays both the barcodes (Figure 2a) and the persistence diagrams (Figure 2b) for all three proteins.

The barcode plots (Figure 2a) show that the vast majority of H_1 and H_2 features are short-lived, corresponding to small-scale topological noise. However, several long-lived features can be observed, repre-

senting more persistent and potentially biologically relevant loops or cavities. In particular, creatine kinase (1CRK) demonstrates a broader range of long-lived H_1 and H_2 features, suggesting greater topological complexity compared to hemoglobin (4HHB) and protein kinase A (1ATP).

The persistence diagrams (Figure 2b) provide an alternative visualization of feature birth and death times, with points near the diagonal representing transient features, and outliers further from the diagonal indicating persistent topological structures. Again, 1CRK stands out for having several H_1 and H_2 features with substantially larger persistence.



(a) Barcodes for H0, H1, and H2.

(b) Persistence diagrams for H0, H1, and H2.

Figure 2: Persistent homology summary for three protein structures: (a) Barcodes; (b) Persistence diagrams for H0, H1, and H2 features of hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP).

To further quantify and visualize the topological complexity of the three representative proteins, we computed Betti curves and persistence entropy for the point cloud data of hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP). The Betti curve, which records the number of topological features (connected components, loops, etc.) as a function of the filtration parameter, provides a comprehensive summary of topological changes across scales. Persistence entropy serves as a summary statistic for the distribution of lifetimes in persistence diagrams, capturing the complexity and diversity of features.

As shown in Figure 3, the Betti curves for all three proteins display a characteristic decrease in Betti numbers as the filtration value increases. This pattern indicates that small-scale features, such as connected components (H0) and loops (H1), are quickly merged or filled in as the filtration progresses. However, subtle differences can be observed: for instance, 1CRK appears to maintain a higher number of persistent loops at intermediate filtration values, suggesting a greater degree of topological complexity and openness in its structure.

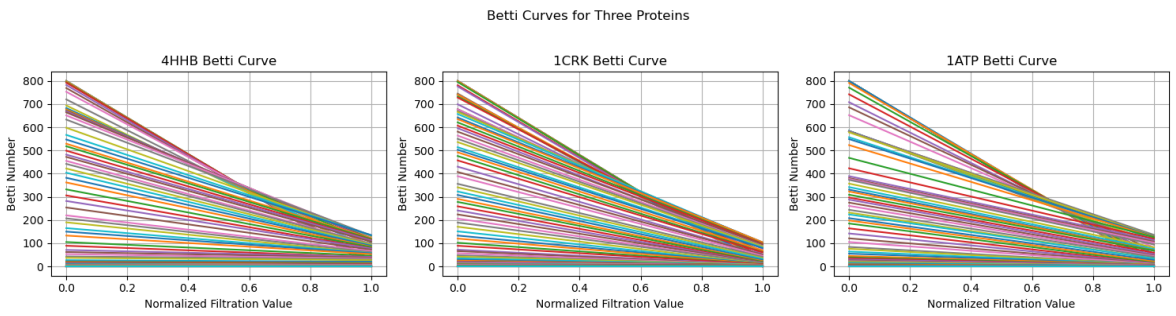


Figure 3: Betti curves for the Vietoris-Rips filtration on the point clouds of hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP). Each curve traces the evolution of Betti numbers across normalized filtration values for sampled batches, summarizing the multiscale topological structure of the proteins.

To complement these qualitative observations, we computed the persistence entropy for both H_0 and H_1 across all proteins. The results are summarized in Table 1.

From Table 1, we observe that all three proteins exhibit similar entropy for H_0, reflecting comparable diversity in their small-scale connected components. In contrast, the entropy values for H_1 suggest modest differences in loop complexity, with protein kinase A (1ATP) displaying the highest H_1 entropy, followed closely by hemoglobin (4HHB). This is consistent with the structural diversity observed in their persistence diagrams and Betti curves. Collectively, the Betti curves and persistence entropy provide complementary perspectives on the multiscale topological organization of protein structures, supporting the identification of subtle yet meaningful differences among them.

Protein	Persistence Entropy H_0	Persistence Entropy H_1
4HHB	9.532	7.817
1CRK	9.544	7.706
1ATP	9.549	7.998

Table 1: Persistence entropy for H_0 (connected components) and H_1 (loops) of each protein. Higher entropy indicates greater diversity and complexity of topological features.

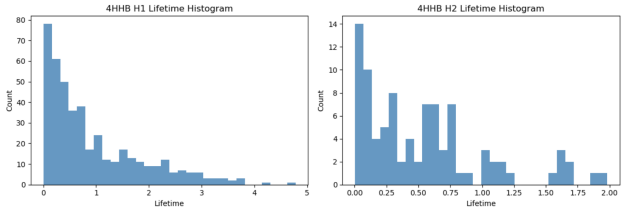
4.3 Quantitative Topological Feature Statistics

Quantitative analysis of the first homology group (H_1) reveals clear differences in the topological complexity of the three protein structures. As summarized in Table 2, hemoglobin (4HHB) contains 439 H_1 loops, creatine kinase (1CRK) has 381, and protein kinase A (1ATP) exhibits 440 loops. Notably, the top three H_1 persistence lifetimes for 1CRK (12.58, 5.81, 5.64) are substantially larger than those for 4HHB (4.78, 4.21, 3.82) or 1ATP (3.45, 3.13, 2.99), indicating that 1CRK possesses more pronounced and persistent loop structures.

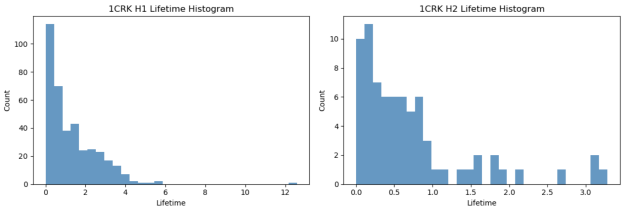
Figure 4 presents the lifetime histograms for H_1 and H_2 features of the three proteins. All distributions are heavily skewed toward short-lived topological features, with a long tail of highly persistent loops and cavities. Notably, creatine kinase (1CRK) shows a greater frequency and range of persistent features, consistent with its more complex structural organization. These findings highlight that persistent homology can distinguish topological noise from long-lived, biologically relevant features.

Protein	H1 Loop Count	Top 1 Persistence	Top 3 Persistences
4HHB	439	4.78	[4.78, 4.21, 3.82]
1CRK	381	12.58	[12.58, 5.81, 5.64]
1ATP	440	3.45	[3.45, 3.13, 2.99]

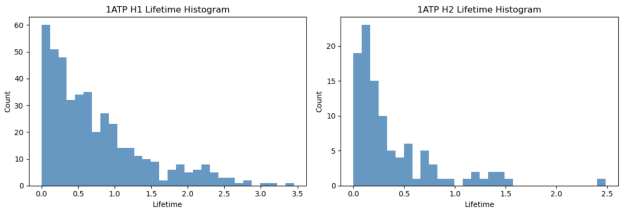
Table 2: Summary of persistent homology analysis for the first homology group (H_1) of three proteins. The table reports the total number of H_1 loops and the three most persistent H_1 features for each structure.



(a) Hemoglobin (4HHB): H_1 and H_2 lifetime histograms



(b) Creatine kinase (1CRK): H_1 and H_2 lifetime histograms



(c) Protein kinase A (1ATP): H_1 and H_2 lifetime histograms

Figure 4: Lifetime histograms of H_1 and H_2 features for each protein. For each structure—hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP)—the distributions are dominated by short-lived topological features, while a small subset displays high persistence. This illustrates both prevalent topological noise and rare, biologically meaningful, long-lived features.

4.4 Spatial Mapping of Persistent Topological Features

To investigate the spatial origin of the most persistent topological features, we mapped the atoms associated with the earliest "birth" radii of the largest H_1 (loop) and H_2 (void) features for each protein structure (Figure 5). For each protein, the grey points represent all non-hydrogen atoms, while the colored points indicate atoms within the minimal spheres responsible for the birth of the top three persistent H_1 or H_2 features. The black cross marks the geometric center of the corresponding birth sphere.

For 1ATP, both the most persistent H_1 and H_2 features originate near the geometric center of the molecule, suggesting the existence of prominent loops and cavities close to the molecular core. In contrast,

for 1CRK, the regions giving rise to the top H_1 and H_2 features are more spatially dispersed, reflecting the more extended and anisotropic architecture of this protein. For 4HHB, persistent features are localized near the center, indicating a relatively compact and globular organization.

This spatial mapping confirms that highly persistent topological features often correspond to structurally significant regions, such as central channels, large cavities, or stable loops within the protein fold. These visualizations further illustrate the distinct spatial organization of the topological fingerprints observed in the three proteins.

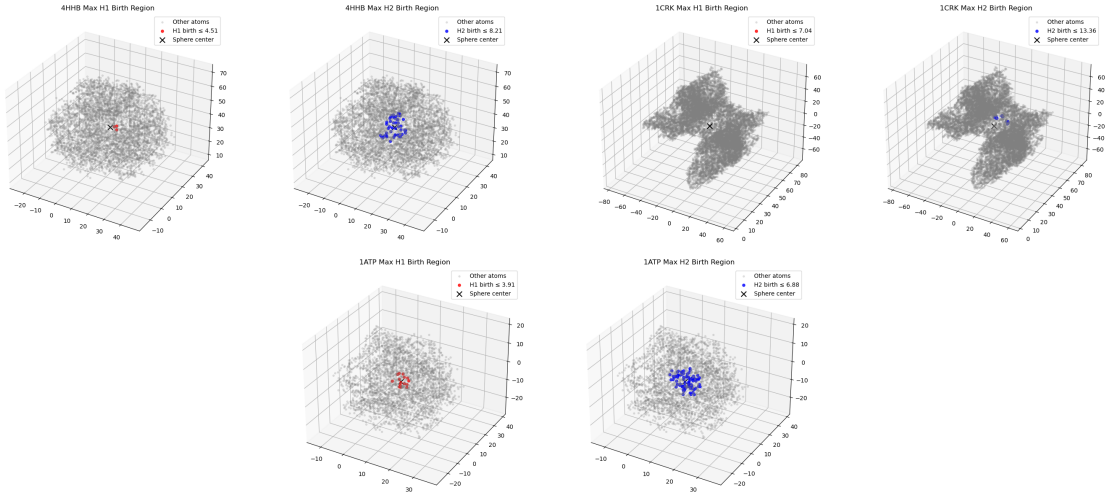


Figure 5: Spatial localization of atoms involved in the formation of the most persistent topological features in each protein: hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP). For each protein, the left panel shows atoms responsible for the top H_1 (loop) features (in red), and the right panel shows those for the top H_2 (void) features (in blue). Grey points indicate all other atoms, and the black cross marks the geometric center of the birth sphere for the maximal feature. These critical regions highlight the structural origin of the most persistent loops and cavities within each protein.

4.5 Inter-Protein Topological Similarity Assessment

Pairwise Wasserstein distances between the H_1 persistence diagrams of the three protein structures are reported in Table 3. The distances are substantial in all cases, with the largest value observed between 1CRK and 1ATP (512.48), followed by 4HHB vs 1CRK (412.35), and 4HHB vs 1ATP (235.92). These results indicate that each protein possesses a distinct topological fingerprint as captured by persistent homology. Notably, creatine kinase (1CRK) is the most topologically dissimilar compared to both hemoglobin (4HHB) and protein kinase A (1ATP), consistent with its more persistent and pronounced H_1 loop features identified earlier. This topological dissimilarity may reflect fundamental differences in molecular organization or biological function among the proteins.

	4HHB	1CRK	1ATP
4HHB	–	412.35	235.92
1CRK	412.35	–	512.48
1ATP	235.92	512.48	–

Table 3: Pairwise Wasserstein distances between the H_1 persistence diagrams of three proteins. Higher values indicate greater topological dissimilarity.

5 Discussion

Our comparative persistent homology analysis of three representative proteins demonstrates the strengths and interpretability of TDA in structural bioinformatics. The Betti curves, persistence diagrams, and entropy measures jointly reveal both shared and distinct topological fingerprints among hemoglobin (4HHB), creatine kinase (1CRK), and protein kinase A (1ATP).

Several key observations emerge. First, while all proteins exhibit abundant short-lived features—reflecting topological noise and small-scale fluctuations—1CRK consistently displays a greater diversity and range of long-lived H_1 and H_2 features. This suggests a more open, anisotropic, and topologically complex fold, potentially linked to unique functional requirements. The localization analysis further highlights that persistent topological features often correspond to the geometric core or extended regions of the protein, indicating their possible biological relevance, such as forming central channels or stable loops that may play a role in ligand binding or conformational flexibility.

The pairwise Wasserstein distances confirm that persistent homology fingerprints can quantitatively discriminate global shape differences beyond traditional geometric or sequence-based descriptors. The particularly high dissimilarity between 1CRK and the other two proteins reflects their marked topological divergence, which may underlie different mechanisms of action or evolutionary origins.

Despite these advances, our study also highlights challenges and limitations. The direct biological interpretation of individual persistent features remains nontrivial, as not every long-lived loop or cavity is necessarily functionally significant. Moreover, choices in filtration, subsampling, and noise handling can influence results. Our pipeline focuses on geometric point clouds, but integrating chemical, evolutionary, or energetic information could further enhance interpretability. Future directions include the systematic association of persistent features with known functional motifs, protein families, or experimental phenotypes, as well as the combination of TDA with machine learning for predictive modeling of structure-function relationships.

Overall, our findings illustrate the utility of persistent homology as a powerful and complementary approach for multiscale protein structure comparison. As TDA tools and biological annotation resources mature, the integration of quantitative topology with functional genomics and drug discovery holds significant promise.

6 Conclusion

In this study, we have applied persistent homology and topological data analysis to the comparative study of protein structures. By extracting and quantifying topological invariants—such as connected components, loops, and cavities—we provide new perspectives on the similarities and differences among hemoglobin, creatine kinase, and protein kinase A.

Our approach highlights the ability of persistent homology to detect subtle yet functionally relevant structural variations that may be overlooked by classical geometric or statistical methods. The integration of Betti curves, entropy statistics, and Wasserstein distances enables a systematic, interpretable, and robust comparison of complex molecular shapes.

While our results reinforce the value of topological approaches for biomolecular structure analysis, they also point to important open questions regarding the biological roles of persistent features and the best practices for integrating topology with traditional descriptors. We anticipate that further developments in TDA, including higher-dimensional analysis, large-scale dataset applications, and the integration with functional and evolutionary information, will continue to advance our understanding of protein structure and function.

In summary, persistent homology provides a mathematically rigorous and biologically insightful frame-

work for protein structure comparison and will be a valuable tool in the expanding repertoire of computational structural biology.

References

[1] Metin E Aktas, Emine E Akbas, Mariusz Jaskolski, and Kelin Xia. Persistence homology of proteins: Theory and practice. *Structure*, 27(3):523–541, 2019.

[2] Mariana Gameiro, Yasuaki Hiraoka, Satoshi Izumi, Konstantin Mischaikow, and Vidit Nanda. Topological measurement of protein compressibility via persistence diagrams. *Japan Journal of Industrial and Applied Mathematics*, 31:399–419, 2014.

[3] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[4] Violeta Kovacev-Nikolic, Peter Bubenik, Dragomir Nikolić, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology*, 15(1):19–38, 2016.

[5] Jerry Tsai, Robin Taylor, Cyrus Chothia, and Mark Gerstein. The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology*, 290:253–266, 1999.

[6] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

[7] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 31(4):e02719, 2015.

[8] Kelin Xia, Zhixiong Zhao, and Guo-Wei Wei. Multiresolution persistent homology for excessively large biomolecular datasets. *Journal of Chemical Physics*, 143(13):134103, 2015.