Using Topological Data Analysis to Compare Hemoglobin Structures

Comparing the topological features of deoxygenated, oxygenated, and sickle-cell hemoglobin.

Madeline Kloberdanz
University of Oregon, Data Science & Biology

Alicia Bierly
University of Oregon, Data Science & Biology

Shannon Raloff
University of Oregon, Mathematics & Statistics

ABSTRACT

Understanding the structural variations of hemoglobin is essential for interpreting its function and dysfunction, especially in the context of diseases such as sickle cell anemia. While traditional structural biology techniques have elucidated many aspects of hemoglobin's conformations, recent advances in Topological Data Analysis (TDA) offer new tools for quantifying subtle changes in protein structure. In this study, we apply persistent homology to compare topological features of three hemoglobin forms, oxygenated, deoxygenated, and sickle-cell hemoglobin, using atomic coordinate data from Protein Data Bank (.pdb) files. We construct Vietoris–Rips complexes and generate persistence diagrams and barcodes to identify structural differences, quantified through bottleneck distances and statistical testing. Results reveal distinguishable topological signatures between oxygenated and deoxygenated hemoglobin, with significant variation in dimension-1 features between sickle-cell and wild-type hemoglobin. These findings suggest that persistent homology can capture meaningful geometric perturbations due to oxygen binding and pathogenic mutation. Our study highlights TDA as a promising computational framework for protein structure analysis, with potential applications in identifying functionally important regions, guiding drug development, and complementing precision medicine approaches.

CCS CONCEPTS • Topology • Hemoglobin • Sickle-cell anemia
Additional Keywords and Phrases: Persistence Diagrams/Barcodes, point clouds, rips complex, simplex trees, machine learning

1 INTRODUCTION

Proteins are three-dimensional molecules whose structure critically determines their function. Hemoglobin, a well-studied example, is responsible for transporting oxygen in the blood and exists in several structural conformations, meaning different spatial arrangements, depending on what is bound to it. Deoxygenated hemoglobin, oxygenated hemoglobin, and the mutated form responsible for sickle-cell anemia each exhibit unique folding patterns that impact their function or biological performance. While traditional biochemical and structural biology approaches have made progress in describing these differences, recent advances in computational topology could offer new tools for quantifying subtle variations in molecular shape and connectivity.

Topological Data Analysis (TDA) is an emerging mathematical framework that studies the "shape" of data through tools such as persistent homology. In the context of molecular biology, TDA allows researchers to

extract topological signatures from protein structures, represented as point clouds derived from atomic 3D spatial coordinates. These signatures, such as persistence diagrams and barcodes, capture multiscale topological features like loops and voids that persist across different scales of resolution. When applied to protein structures, these features can reflect changes in binding site geometry or structural flexibility.

In this project, we apply persistent homology to compare the topological features of three types of hemoglobin: oxygenated hemoglobin, deoxygenated hemoglobin, and sickle-cell hemoglobin. Using the Gudhi library in Python, we constructed Vietoris–Rips complexes from atomic coordinate data extracted from .pdb files and computed persistence diagrams to analyze the resulting topological features. To compare these structures quantitatively, we also calculated bottleneck distances and performed t-tests between persistence diagrams.

Our approach is inspired by recent work applying TDA to protein flexibility and binding affinity [1, 2], as well as the broader promise of TDA in biological shape analysis [3]. We aim to explore how a topological perspective can illuminate the effects of mutation and conformational change on protein function. In particular, we focus on how the structural deformation caused by the Glu6Val mutation in sickle-cell hemoglobin manifests in topological space and whether such changes can be detected through persistent homology.

2 HEMOGLOBIN STRUCTURE

Hemoglobin is a tetrameric protein complex responsible for oxygen transport in vertebrate blood. It is composed of four globular subunits: two alpha (α) chains and two beta (β) chains. There are several forms of hemoglobin which are denoted by their confirmation, meaning changes in the structure which changes its function. Each subunit contains a heme group consisting of a porphyrin ring with an iron atom in the center which is capable of binding one oxygen molecule.

Oxygenated hemoglobin is the form in which all four heme groups have bound oxygen molecules. In this conformation, the protein adopts a more compact and symmetrical structure. Binding of oxygen induces subtle shifts in the relative positions of the subunits. These shifts affect the surface structure.

Deoxygenated hemoglobin (the "tense" or T-state) lacks bound oxygen and exhibits a different quaternary structure from its oxygenated form. These conformational changes involve shifts in the orientation of helices and atomic coordinates.

Sickle cell hemoglobin (HbS) is a genetic variant of adult hemoglobin, in which a mutation creates a hydrophobic patch on the surface of the β-subunit, creating small differences in surface chemistry can lead to large-scale topological differences during polymerization.

3 PERSISTENT HOMOLOGY

A filtration is a sequence of subspaces of the topological space connected by inclusions, and persistent homology is a tool used to describe the holes in topological spaces formed by nontrivial cycles in these sublevel sets of the topological space. This can be used as a technique in topological data analysis since the interval between the birth and death of features can be used to create something called a 'persistence barcode' or 'persistence diagram' which is a description of the topological space. This form of data analysis can capture the global shape of complex data and is less susceptible to local perturbations. The

length of persistence of a feature can be a form of 'importance' and allows the comparison of different persistence diagrams through the definition of penalties for distance between two persistence profiles.

A simplex is the convex hull of a set of independent points, and a k-simplex is defined to have dimension k formed from a collection of k+1 points. A simplicial complex K is a collection of finitely many simplices that contains every face of each simplex in K, and its dimension is the maximum dimension of any simplex in the simplicial complex. Simplicial complexes can be formed from open covers of point clouds using the concept of 'nerves' which is the contraction of nonempty intersections of cover elements of a metric space to points. This concept can be used to create a Cech complex, which is the homotopy equivalent to the space of the union of balls of a certain radius covering a point cloud. The Cech complex is then related to another complex called the Vietoris-Rips complex, which has less computational burden and is roughly equivalent approximation of the more theoretically rigorous Cech complex, as the inclusion of simplices in a filtration interleave at similar alpha values of radii for the balls of the point cloud.

4 METHODS

We analyzed the 3D structural topology of hemoglobin structures using publicly available data from the RCSB Protein Data Bank (RCSB PDB) and parsed the structure files using Bio.PDB and Bio.PDB.MMCIFParser modules from Biopython to extract relevant information including residue names and numbers, atom names, elements, and three-dimensional coordinates (x, y, z). To reduce complexity, we calculated the mean of the atomic coordinates within each residue, yielding a single representative point. These mean coordinates were used to construct point clouds representing the spatial configuration.

To investigate topological features, we used the Gudhi library to generate Vietoris-Rips complex complexes from the point clouds with a maximum edge length of 25. Simplex trees were constructed to dimension 2 to capture the connected components (dimension 0), loops/holes (dimension 1), and cavities (dimension 2) in these structures. From these, we created persistence diagrams and barcode plots with a minimum persistence threshold of 1 to filter out noise in the data.

To assess structural variation, bottleneck distances were computed between persistence diagrams across multiple structures and dimensions. In addition, independent t-tests were conducted on the birth and death times of features in each dimension to determine statistical significant differences.

5 RESULTS

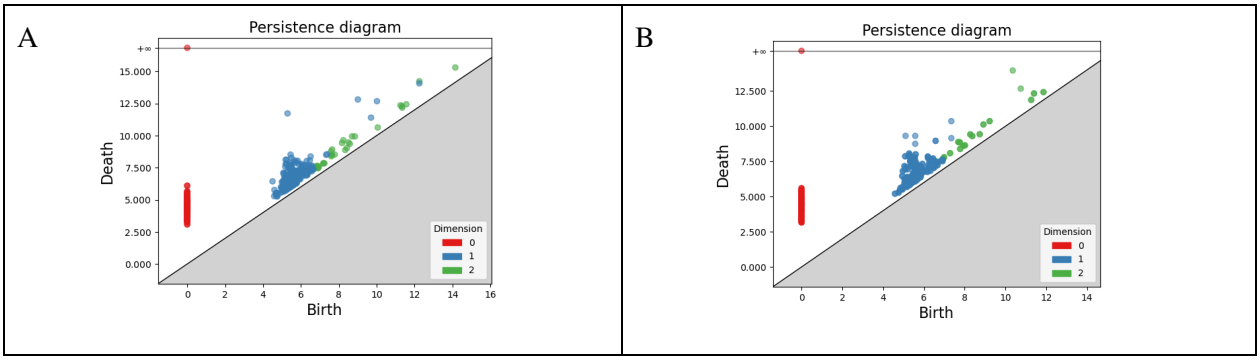

Figure 1. Persistence Diagrams for deoxygenated hemoglobin (A) and oxygenated hemoglobin (B).

Figure 1 shows the persistence diagrams for deoxygenated hemoglobin (A) and oxygenated hemoglobin at all four active sites (B). The most notable differences appear between the two structures in the birth and deaths of dimensions 1 and 2. Bottleneck distances between the corresponding diagrams were computed as 1.579 for dimension 0, 2.0516 for dimension 1, and 1.5236 for dimension 2, suggesting moderate structural differences across these homological dimensions. However, independent *t*-tests comparing the birth and death values across all three dimensions revealed no statistically significant differences between the oxygenated and deoxygenated forms.

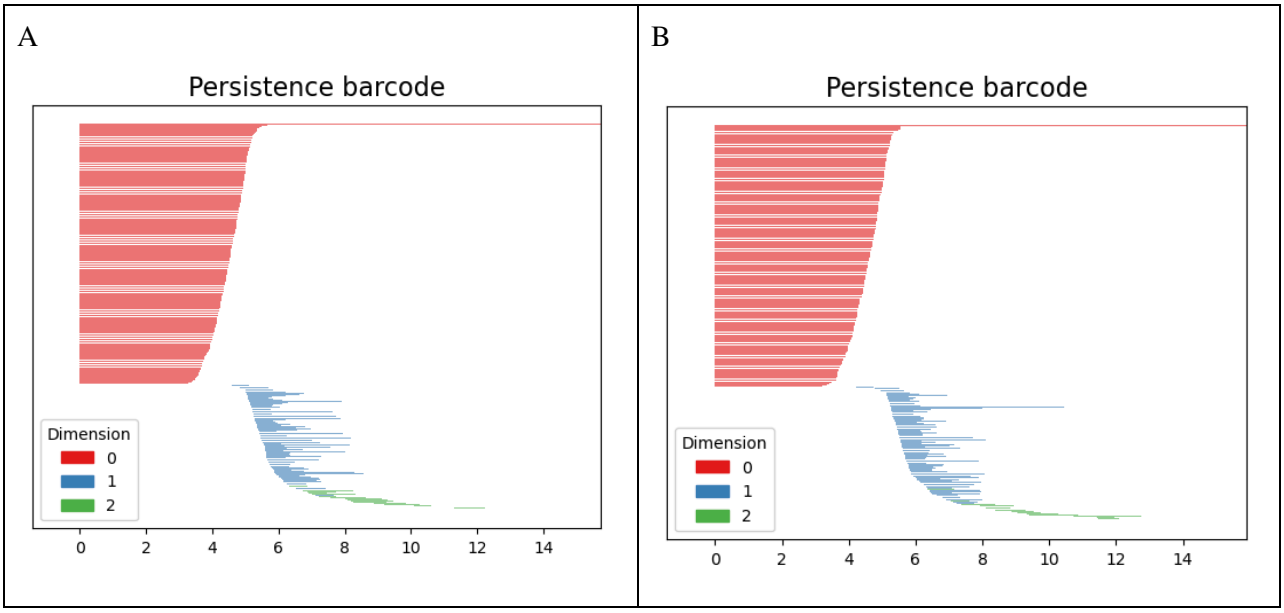

Figure 2. Persistence barcode of sickle cell hemoglobin (A) and wildtype hemoglobin (B)

Figure 2 shows the persistence barcodes with significant features extracted from all three dimensions of sickle cell hemoglobin (A) and wild type hemoglobin (B). Independent t-tests revealed significant differences in birth rates of dimension 1 features, with sickle cell hemoglobin showing a higher average birth scale ($t = 1.999$, $p = 0.0459$, $df = 922$). Bottleneck distances between the persistence diagrams of the two hemoglobin were calculated, with distances of 1.645 for dimension 0, 1.493 for dimension 1, and 0.821 for dimension 2. When beta subunits were isolated, similar values were found for dimensions 0 and 1, but dimension 2 decreased to a value of 0.419.

5 DISCUSSION

5.1 Interpretation of Results

The primary differences between oxygenated and deoxygenated hemoglobin are most evident in dimensions 1 and 2 of the persistence diagrams. Oxygenated hemoglobin displays more tightly clustered organized features in dimension 1, suggesting great uniformity in one-dimensional holes. This aligns with biochemical understanding that oxygen binding promotes the relaxed state of hemoglobin, which enhances cooperative interactions and binding between subunits and oxygen which stabilizes the overall structure.

The bottleneck distances support the structural conformation differences as they have relatively high values in all three dimensions. The differences in these structures are due to rotation of 15° of a $\alpha\beta$ dimer,

shift of 0.8 Å by a $\alpha\beta$ dimer, and increased compactness in oxygenated form. These changes affect the overall structure and orientation of coordinates in the three dimensional space. The higher bottleneck distance in dimension 1 suggests that this confirmation change affects loops or voids more than it does in connected components or higher dimensional cavities. Oxygenation may reduce or eliminate one-dimensional features due to compactness.

Comparing sickle cell hemoglobin and wild-type hemoglobin has different patterns. A significant difference was found in the birth values of dimension 1 features, with sickle cell hemoglobin exhibiting higher average birth values. This means there is a shift at which one-dimensional holes emerge from the structure, likely showing structural differences in the beta subunits as of a result of the hydrophobic patch.

In addition, it is interesting we got similar values for the bottleneck distances for dimension 0 and 1 using the entire point cloud vs only the beta subunits when looking at sickle cell and wild-type hemoglobin. This supports the idea that TDA could be used on targeted regions of a protein, rather than requiring analysis of the entire molecular structure. This flexibility could be used for high-resolution structural data or when researchers want to isolate the effects of specific mutations, binding sites, or allosteric changes.

5.2 Limitations of Analysis

Although independent t-tests were performed between the birth and death rates of sickle cell, deoxygenated, and oxygenated hemoglobin most didn't have significance. This may be partly attributed to computational limitations that required us to cap the maximum filtration length to 25, resulting in the simplex tree explaining around 30% of the possible distance values, restricting longer-lived topological features.

In addition, our analysis included only a single structure for each hemoglobin variant. This limited our ability to draw generalizable or statistically supported conclusions and prevented us from applying newly researched techniques. Expanding to include multiple structures per hemoglobin type, whether from different conformational states or patient-derived samples, would enable more robust analysis. Statistical methods such as permutation testing or bootstrapping could assess the significance of observed topological differences [1]. Group-level comparisons using bottleneck or Wasserstein distances across persistence diagrams could reveal consistent topological trends tied to biological properties, such as oxygen-binding efficiency or the propensity for aggregation.

Another limitation is the use of atomic coordinates without associated atomic radii, which restricts the realism of our point cloud models. This detail is especially important, considering how important atomic size is in intramolecular interactions. Incorporating this information could better represent steric constraints and more accurately model conformational changes. Additionally, scaling to larger datasets presents computational challenges, particularly in generating Rips complexes and computing simplex trees, which would require optimized algorithms or high-performance computing resources.

6 AREAS OF FURTHER EXPLORATION

6.1 Connecting TDA to Broader Applications in Protein Biology

Our analysis revealed that persistent homology can capture subtle structural differences between oxygenated, deoxygenated, and sickle cell hemoglobin. These topological features offer a unique lens for

interpreting protein structure beyond traditional metrics. In future studies, these features could be linked to biologically relevant properties such as binding affinity, cooperative oxygen loading, or stability. One study by Kovacev-Nikolic, et al. demonstrates that topological tools can localize functionally important regions in proteins. Extending this logic, persistent homology may help identify allosteric sites or regions that differ structurally between functional and dysfunctional protein structures [1].

A study by Bramer and Wei approached looking at protein structures using Atom-Specific Persistent Homology by using element specific subsets of the structures in order to predict protein B-factors (a measure of atomic thermal motion) [2]. This would reduce the complexity of the point clouds and also allow for chemical specificity of the data which could then be applied to see changes in conformations more specifically. In addition, this study included 60 total topological features which included a combination of global features (protein resolution and r-value) and local features (secondary structure and packing density). The study then used gradient boosted trees and convolutional neural networks to predict the B-factors.

6.3 Antisickling Agents and Topological Insights

While breakthrough treatments like CRISPR-based gene editing (e.g., Casgevy) have offered the promise of a functional cure for sickle cell anemia, their implementation remains limited. With treatment costs exceeding $3 million and eligibility restrictions due to clinical or genetic criteria, these approaches are currently out of reach for many patients. Therefore, it remains essential to explore more accessible and widely applicable therapeutic strategies. Antisickling agents, such as clotrimazole and hydroxyurea, aim to reduce HbS polymerization or increase fetal hemoglobin levels, improving symptoms and quality of life [4].

This therapeutic diversity opens a new area of inquiry for Topological Data Analysis (TDA). Can persistent homology detect structural shifts in hemoglobin before and after drug binding or HbF induction? Could TDA be used to compare the conformational landscape of hemoglobin under different therapeutic conditions, such as with and without hydroxyurea treatment? These questions motivate future research into whether topological signatures can be used to predict or assess the efficacy of antisickling agents, providing insight into their mechanisms and guiding the development of new compounds.

6.4 Future Integration of TDA in Biomedical Research

Topological Data Analysis offers a complementary framework to traditional structure-based analyses. By capturing global shape and connectivity information, TDA can highlight differences that may be functionally relevant but subtle or non-local. In the future, TDA could aid in the identification of biomarkers, or help prioritize protein conformations for drug targeting. Furthermore, integration with machine learning could enable classification of pathogenic versus benign mutations based on topological descriptors, broadening the utility of TDA in structural biology and precision medicine.

7 BIBLIOGRAPHY

[1] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. 2016. Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology* 15, 1 (2016), 19–38. https://doi.org/10.1515/sagmb-2015-0057

[2] David Bramer and Guo-Wei Wei. 2020. Atom-specific persistent homology and its application to protein flexibility analysis. *Computational and Mathematical Biophysics* 8, 1 (2020), 1–17. https://doi.org/10.1515/cmb-2020-0001

[3] M. Dindin, Y. Umeda, and F. Chazal. 2020. Topological Data Analysis for Arrhythmia Detection Through Modular Neural Networks. In *Proceedings of the 33rd Canadian Conference on Artificial Intelligence (Canadian AI 2020)*, Lecture Notes in Computer Science, Vol. 12109. Springer, Cham, 141–152. https://doi.org/10.1007/978-3-030-47358-7_17

[4] Antisickling Agent – an overview | ScienceDirect Topics. Retrieved June 10, 2025 from https://www.sciencedirect.com/topics/medicine-and-dentistry/antisickling-agent