Received: 14 May 2017 | Revised: 27 June 2017 | Accepted: 29 June 2017

DOI: 10.1002/cnm.2914

RESEARCH ARTICLE - FUNDAMENTAL

WILEY

Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction

Zixuan Cang[1] | Guo-Wei Wei[1,2,3]

[1] Department of Mathematics, Michigan State University, MI 48824, USA

[2] Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

[3] Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA

Correspondence
Guo-Wei Wei, Department of Mathematics Michigan State University, MI 48824, USA.
Email: wei@math.msu.edu

Funding information
National Science Foundation, Grant/Award Number: Division of Information and Intelligent Systems, 1302285 and Directorate for Mathematical Sciences, 1721024, ; MSU Center for Mathematical Molecular Biosciences Initiative

Abstract

Protein-ligand binding is a fundamental biological process that is paramount to many other biological processes, such as signal transduction, metabolic pathways, enzyme construction, cell secretion, and gene expression. Accurate prediction of protein-ligand binding affinities is vital to rational drug design and the understanding of protein-ligand binding and binding induced function. Existing binding affinity prediction methods are inundated with geometric detail and involve excessively high dimensions, which undermines their predictive power for massive binding data. Topology provides the ultimate level of abstraction and thus incurs too much reduction in geometric information. Persistent homology embeds geometric information into topological invariants and bridges the gap between complex geometry and abstract topology. However, it oversimplifies biological information. This work introduces element specific persistent homology (ESPH) or multicomponent persistent homology to retain crucial biological information during topological simplification. The combination of ESPH and machine learning gives rise to a powerful paradigm for macromolecular analysis. Tests on 2 large data sets indicate that the proposed topology-based machine-learning paradigm outperforms other existing methods in protein-ligand binding affinity predictions. ESPH reveals protein-ligand binding mechanism that can not be attained from other conventional techniques. The present approach reveals that protein-ligand hydrophobic interactions are extended to 40Å away from the binding site, which has a significant ramification to drug and protein design.

KEYWORDS

protein-ligand binding affinity, machine learning, topology

1 | INTRODUCTION

The study of protein-ligand binding has attracted enormous attention in the past few decades due to its importance to biochemistry, biophysics, and biomedicine.[1-3] Experimentally, 3-dimensional (3D) structures of protein-ligand complexes obtained from X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance shed light on molecular recognition, protein conformational changes upon binding, and possible allosteric effect. Sophisticated physical and chemical tools, including quantum mechanical calculation, molecular dynamics simulation, and Monte Carlo sampling have been applied to protein-ligand binding analysis. Nevertheless, despite of much effort, our ability to predict binding affinities is still quite limited, which suggests a significant gap in our understanding.

Essentially, there are 3 types of methods for protein-ligand binding predictions: physics based,[4,5] knowledge based,[6,7] and empirical ones.[8-11] In general, physics-based methods invoke molecular mechanism and/or implicit solvent approaches to provide unique insights into the molecular mechanism of protein-ligand interactions. A prevalent view is that binding involves intermolecular forces, such as steric contacts, ionic bonds, hydrogen bonds, hydrophobic effects, and van der Waals interactions. Meanwhile, empirical methods might work well with carefully selected data sets. However, both physics-based methods and empirical methods are not designed to deal with increasingly diverse and rapidly growing data sets. The application of accurate physical methods in the circumstances with large data sets is limited by the demand of huge amount of computing resources, and it is challenging to apply empirical methods to increasingly diverse data sets. Additionally, most data sets of protein-ligand complexes generated from x-ray crystallography have limited resolutions around 2 Å and typically do not include hydrogen atoms. NMR structures have less resolutions in general. For this type of data sets, physics-based models often contain too much geometric detail that are not available from experimental data. Knowledge-based methods resort to modern machine-learning techniques, which use nonlinear regressions and exploit large data sets to uncover hidden patterns in data sets and are able to outperform other methods in massive and complex data challenges. However, data-driven feature selections in knowledge-based scoring functions tend to abandon physical consideration and render a high-dimensional problem.

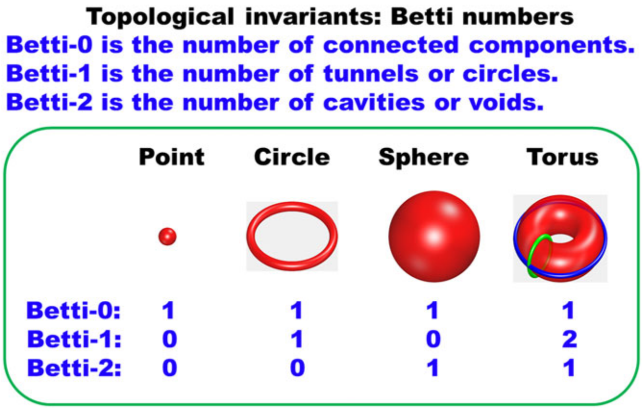
In this work, we propose an entirely new strategy that integrates persistent homology[12,13] and machine learning to elucidate molecular mechanism in protein-ligand binding and predict binding affinities. Mathematically, most existing methods are structure or geometry-based models that are often inundated with too much structural detail resting in excessively high dimensions. In contrast, topology deals with the connectivity of different components in a space and characterizes independent entities, rings, and higher dimensional topological faces within the space.[14] It provides the ultimate level of abstraction of many biological processes, such as the open or close of ion channels, the assembly or disassembly of virus capsids, the folding and unfolding of proteins, and the association or dissociation of ligands.[13,15-21] However, conventional topology or homology is truly free of metrics or coordinates and thus retains too little geometric information to be practically useful. Persistent homology is a new branch of algebraic topology that embeds multiscale geometric information into topological invariants to achieve the interplay between geometry and topology.[22-31] Efficient computational algorithms have been developed for persistent homology.[32-37] Nevertheless, the direct application of persistent homology to macromolecules without considering atomic properties oversimplifies biological information and thus is not very useful in quantitative predictions. We introduce element specific persistent homology (ESPH) to dramatically reduce biomolecular complexity while retaining crucial chemical and biological information. ESPH offers a multicomponent persistent homology representation and enables us to decipher the entangling code of protein-ligand interactions. We also propose interactive persistent homology (IPH) to describe interactions between protein and ligand. Finally, we develop binned barcode representation (BBR) to characterize the strength of various protein-ligand interactions. The proposed ESPH, IPH, and BBR give rise to an appropriate level of topological abstraction, geometric specificity, and biological information to reveal the molecular mechanism of protein-ligand binding and deliver a concise, efficient, accurate while low dimensional representation of protein-ligand interactions. Barcodes generated by the persistent homology computation are called topological fingerprints (TF)[38] or element specific topological fingerprints (ESTF) if ESPH is used. Features to be fed to machine-learning methods are generated from TFs/ESTFs using various treatments including BBR. Our topology-based binding prediction method (T-Bind) is validated on 2 large benchmark data sets, and it outperforms the other methods whose benchmark results are available on the same data sets.

2 | METHODS AND ALGORITHMS

2.1 | Persistent homology theory

2.1.1 | Overview

The fundamental task of topological data analysis is to extract topological invariants, namely, the intrinsic features of the underlying space, of a given data set without additional structure information, like covalent bonds, hydrogen bonds, and van der Waals interactions. A fundamental concept in algebraic topology is simplicial homology, which concerns the identification of topological invariants from a set of discrete nodes such as atomic coordinates in a protein-ligand complex.



Topological invariants: Betti numbers

Betti-0 is the number of connected components.

Betti-1 is the number of tunnels or circles.

Betti-2 is the number of cavities or voids.

	Point	Circle	Sphere	Torus
Betti-0:	1	1	1	1
Betti-1:	0	1	0	2
Betti-2:	0	0	1	1

FIGURE 1 An illustration of topological invariants, ie, Betti numbers for a point, a circle, an empty sphere, and a torus. For the torus, 2 auxiliary rings are added to explain Betti-1= 2

For a given configuration, independent components, rings, and cavities are topological invariants and their numbers are called Betti-0, Betti-1, and Betti-2, respectively; see Figure 1. To study topological invariants in a discrete data set, simplicial homology uses a specific rule such as Vietoris-Rips (VR) complex, Cĕch complex or alpha complex to identify simplicial complexes from simplexes. A 0-simplex is vertex, a 1-simplex an edge, a 2-simplex a triangle, and a 3-simplex represents a tetrahedron. Algebraic groups built on simplicial complexes are used in simplicial homology to systematically compute various Betti numbers. The basic concepts of persistent homology, including simplex, simplicial complex, homology, and filtration, are reviewed below. Two types of simplicial complexes, Vietoris-Rips complex and alpha complex, used in this work are also described.

2.1.2 | Simplex and simplicial complex

A k-simplex is a convex hull of $k+1$ vertices, which is represented by a set of affinely independent points

$$\sigma = \{\lambda_0 u_0 + \lambda_1 u_1 + \dots + \lambda_k u_k | \sum \lambda_i = 1, \lambda_i \geqslant 0, i = 0, 1, \dots, k\}, \tag{1}$$

where $\{u_0, u_1, \dots, u_k\} \subset \mathbb{R}^k$ is the set of points, σ is the k-simplex, and constraints on λ_i's ensure the formation of a convex hull. A convex combination of points can have at most $k+1$ points in $\mathbb{R}^k$. For example, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and a 4-simplex is a 5-cell. A subset of the $k+1$ vertices of a k-simplex with $m+1$ vertices forms a convex hull in a lower dimension and is called an m-face of the k-simplex. An m-face is proper if $m < k$. The boundary of a k-simplex σ is defined as a formal sum of all its $(k-1)$-faces as

$$\partial_k \sigma = \sum_{i=0}^{k} [u_0, \dots, \hat{u}_i, \dots, u_k]^k (-1)^i [u_0, \dots, \hat{u}_i, \dots, u_k], \tag{2}$$

where $[u_0, \dots, \hat{u}_i, \dots u_k]$ denotes the convex hull formed by vertices of σ with the vertex u_i excluded and ∂_k is called boundary operator.

A collection of finitely many simplices forms a simplicial complex denoted by $\mathcal{K}$ satisfying following conditions.

- Faces of any simplex in $\mathcal{K}$ are also simplices in $\mathcal{K}$.
- Intersection of any 2 simplices can only be face of both or an empty set.

2.1.3 | Homology

Given a simplicial complex $\mathcal{K}$, a k-chain c_k of $\mathcal{K}$ is a formal sum of the k-simplices in $\mathcal{K}$ with k no greater than dimension of $\mathcal{K}$ and is defined as $c_k = \sum a_i \sigma_i$ where σ_i is the k-simplices and a_i is coefficients. Generally, a_i can be set within different fields such as $\mathbb{R}$ and $\mathbb{Q}$, or integers $\mathbb{Z}$. For simplicity, a_i is chosen to be $\mathbb{Z}_2$, which is most widely used in computational topology. Let the group of k-chains in $\mathcal{K}$ be denoted by C_k. C_k with addition operation of modulo 2 addition then form an Abelian group $(C_k, \mathbb{Z}_2)$. We now can extend the definition of the boundary operator introduced in Equation 2 to chains.

The boundary operator applied to a k-chain c_k is defined as

$$\partial_k c_k = \sum a_i \partial_k \sigma_i, \tag{3}$$

where σ_is are k-simplices. Therefore, the boundary operator is a map from C_k to C_{k-1}, which is also named boundary map for chains. Note that operator ∂_k satisfies the property that $\partial_k \circ \partial_{k+1} \sigma = 0$ for any $(k+1)$-simplex σ following the fact that any $(k-1)$-face of σ is contained in exactly 2 k-faces of σ. The chain complex is defined as a sequence of chains connected by boundary maps with an order of decaying in dimensions and is represented as

$$\cdots \to C_n(\mathcal{K}) \xrightarrow{\partial_n} C_{n-1}(\mathcal{K}) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(\mathcal{K}) \xrightarrow{\partial_0} 0. \tag{4}$$

The k-cycle group and k-boundary group are defined as kernel and image of ∂_k and ∂_{k+1}, respectively,

$$\begin{aligned} \mathcal{Z}_k &= \mathrm{Ker}\partial_k = \{c \in C_k | \partial_k c = 0\}, \\ \mathcal{B}_k &= \mathrm{Im}\partial_{k+1} = \{\partial_{k+1} c | c \in C_{k+1}\}, \end{aligned} \tag{5}$$

where $\mathcal{Z}_k$ is the k-cycle group and $\mathcal{B}_k$ is the k-boundary group. Since $\partial_k \circ \partial_{k+1} = 0$, we have $\mathcal{B}_k \subseteq \mathcal{Z}_k \subseteq C_k$. With the aforementioned definitions, the k-homology group is defined to be the quotient group by taking k-cycle group modulo of k-boundary group as

$$\mathcal{H}_k = \mathcal{Z}_k / \mathcal{B}_k, \tag{6}$$

where $\mathcal{H}_k$ is the k-homology group. The kth Betti number is defined to be rank of the k-homology group as $\beta_k = rank(\mathcal{H}_k)$.

2.1.4 | Filtration and persistent homology

For a simplicial complex $\mathcal{K}$, we define a filtration of $\mathcal{K}$ as a nested sequence of subcomplexes of $\mathcal{K}$,

$$\emptyset = \mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \cdots \subseteq \mathcal{K}_n = \mathcal{K}. \tag{7}$$

In persistent homology, the nested sequence of subcomplexes usually depends on a filtration parameter. The homology of each subcomplex is analyzed and the persistence of a topological feature is represented by its life span with respect to filtration parameter. Subcomplexes corresponding to various filtration parameters offer the topological fingerprints of multiple scales. The kth persistent Betti numbers $\beta_k^{i,j}$ are ranks of kth homology groups of $\mathcal{K}_i$, which are still alive at $\mathcal{K}_j$ and are defined as

$$\beta_k^{i,j} = \mathrm{rank}(\mathcal{H}_k^{i,j}) = \mathrm{rank}(\mathcal{Z}_k(\mathcal{K}_i)/(\mathcal{B}_k(\mathcal{K}_j) \cap \mathcal{Z}_k(\mathcal{K}_i))). \tag{8}$$

These persistent Betti numbers are used to represent topological fingerprints with their persistence.

2.1.5 | VR complex

Given a metric space M and a cutoff distance d, a simplex is formed if all points in it has pairwise distances no greater than d and all such simplices form the VR complex. The abstract property of the Vietoris-Rips complex enables the construction of simplicial complexes for correlation function-based metric spaces, which models pairwise interaction of atoms with correlation functions instead of native spatial metrics.

2.1.6 | Alpha complex

While Vietoris-Rips complex falls into the category of abstract simplicial complexes, alpha complex provides geometric realization. Given a finite point set X in $\mathbb{R}^n$, a Voronoi cell for a point x is defined as

$$V(x) = \{y \in \mathbb{R}^n | |y - x| \leqslant |y - x'|, \forall x' \in X\}. \tag{9}$$

Given an index set I and a corresponding collection of open sets $\mathbf{U} = \{U_i\}_{i \in I}$, which is a cover of points in X, the nerve of $\mathbf{U}$ is defined as $\mathbf{N}(\mathbf{U}) = \{J \subseteq I | \cap_{j \in J} U_j \neq \emptyset\} \cup \emptyset$. A nerve is an abstract simplicial complex. When the cover $\mathbf{U}$ of X is constructed by assigning a ball of given radius δ, the corresponding nerve forms the simplicial complex named Čech complex,

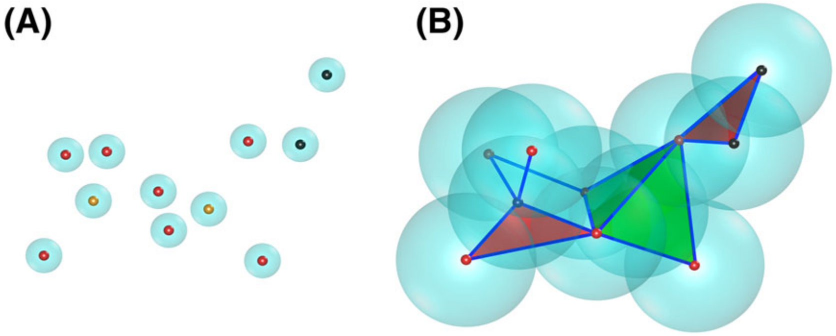


FIGURE 2 An illustration of simplicial complex and filtration of ligand heavy atoms in protein 3LPL. Red balls are oxygen atoms, black balls are carbon atoms, and orange balls are phosphorus atoms. A, Positions of heavy atom spheres are generated via the radius filtration at $r = 0.7$Å. In this case, one has eleven 0-simplexes, zero 1-simplex and zero 2-simplex. B, Filtration progresses to $r = 2.56$Å. In this case, one has zero 0-simplex, three 1-simplex, two 2-simplexes, and one 3-simplex

$$C(X,\delta) = \{\sigma | \cap_{x\in\sigma} B(x,\delta) \neq \emptyset\}, \tag{10}$$

where $B(x,\delta)$ is a closed ball in $\mathbb{R}^n$ with x as the center and δ as the radius. The alpha complex is constructed with cover of X, which contains intersection of Voronoi cells and balls,

$$A(X,\delta) = \{\sigma | \cap_{x\in\sigma} (V(x) \cap B(x,\delta)) \neq \emptyset\}. \tag{11}$$

In this work, VR complex is applied with various correlation-based metric spaces to analyze pairwise interaction patterns between atoms and possibly extract abstract patterns of interactions while alpha complex is applied with Euclidean space of $\mathbb{R}^3$ to identify geometric features such as voids and cycles which may play a role in regulating protein-ligand binding processes. The software packages Dipha[37] and Dionysus[39] are used in this work.

2.2 | Persistent homology analysis of protein-ligand complexes

Simplicial homology is metric free and thus is too abstract to be insightful for complex and large protein-ligand binding data sets. Persistent homology consists of a series of homologies constructed over a filtration process, in which the connectivity of the given data set is systematically reset according to a scale parameter. In the Euclidean distance–based filtration for biomolecular coordinates, the scale parameter is an ever-increasing radius of an ever-growing ball whose center is the coordinate of each atom; see Figure 2. Therefore, filtration-induced persistent homology gives a multiscale representation of the corresponding topological space and reveals topological persistence of the given data set.

The power of persistent homology lies in its topological abstraction and dimensionality reduction. It reveals topological connectivity in biomolecular complexes in terms of TFs,[40-43] which are recorded as the barcodes[44] of biomolecular topological invariants over filtration. It is worthy to mention that topological connectivity differs from chemical bonds, van der Waals bonds or hydrogen bonds. Indeed, TFs offer an entirely new representation of protein-ligand interactions. Figure 3 depicts the TFs of protein-ligand complex 3LPL. By a comparison of TFs of the protein and those of the corresponding protein-ligand complex near the binding site, changes in Betti-0, Betti-1, and Betti-2 panels can be easily noticed. For example, more bars occur in the Betti-1 panel around filtration parameter values 3Å to 5Å after the binding, which indicates a potential hydrogen bonding network due to protein-ligand binding. Additionally, binding induced Betti-2 bars in the range of 4Å to 6Å reflect potential protein-ligand hydrophobic contacts. In fact, changes in Betti-0 bars are associated with ligand atomic types and atomic numbers. Therefore, TFs and their changes describe protein-ligand binding in terms of topological invariants.

2.3 | Element specific persistent homology analysis of protein-ligand complexes

However, these collective TFs ignore crucial biological information and have a limited power in characterizing biomolecules. To characterize biomolecular systems, we introduce ESPH. Specifically, we consider commonly occurring heavy element types in a protein-ligand complex, namely, C,N,O, and S in proteins and C, N, O, S, P, F, Cl, Br, and I in ligands. Our ESPH reduces biomolecular complexity by disregarding individual atomic character, while retaining vital biological information by distinguishing element types. Additionally, to characterize protein-ligand interactions,

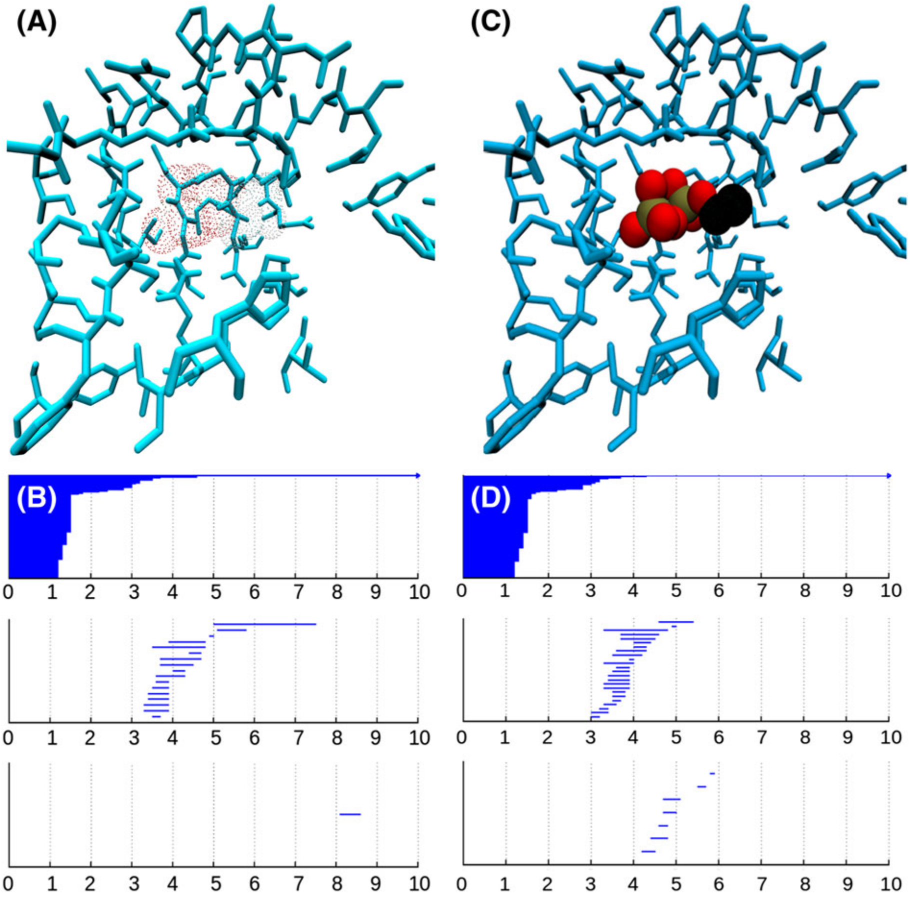


FIGURE 3 Illustration of protein-ligand binding induced topological fingerprints (TF) change. Hydrogen atoms are shown but not used in TF computation. The unit is Å for the barcode plots. A, Binding site residues of protein 3LPL. B, The TFs of the heavy atoms of protein 3LPL binding site residues without the ligand. C, Binding site residues of protein 3LPL and the ligand. D, The TFs of the heavy atoms of protein 3LPL binding site residues and the ligand

we introduce IPH by selecting a set of heavy atoms involving a pair of element types, one from protein and the other from ligand, within a given cutoff distance. The resulting TFs, called interactive ESTFs, are able to characterize intricate protein-ligand interactions. For example, interactive ESTFs between oxygen atoms in the protein and nitrogen atoms in the ligand unveil possible hydrogen bonds, while interactive ESTFs from protein carbon atoms and ligand carbon atoms indicate hydrophobic effects as shown in Figure 4. The detailed construction of TFs/ESTFs and resulting features are described in the following section.

2.4 | Topological fingerprint and feature construction

2.4.1 | Correlation functions

When modeling 3-D structure of proteins, interactions between atoms are related to spatial distances and atomic properties. However, Euclidean metric space does not directly give quantitative description of interaction strengths of atomic interactions. A nonlinear function is applied to map the Euclidean distances together with atomic properties to a measurement of correlation or interaction between atoms. Computed atomic pairwise correlation values form a correlation matrix that can then be used to analyze connectivity patterns between clusters of atoms. In the rest of this section, functions that map geometric distance to topological connectivity are referred to as kernels.

In our previous work, a flexibility-rigidity index (FRI) theory[45] is introduced, which uses decaying radial basis functions to quantify pairwise atomic interactions or correlations. The correlation matrix is then applied to analyze flexibility and rigidity of the protein. The flexibility index computed from the correlation matrix has been found to strongly correlate to experimental B-factors. It has also found its success in the prediction of protein motion[46] and the modeling of fullerene stability.[47] In previous studies, the most favorable outcome is obtained with mainly 2 types of kernels, the exponential

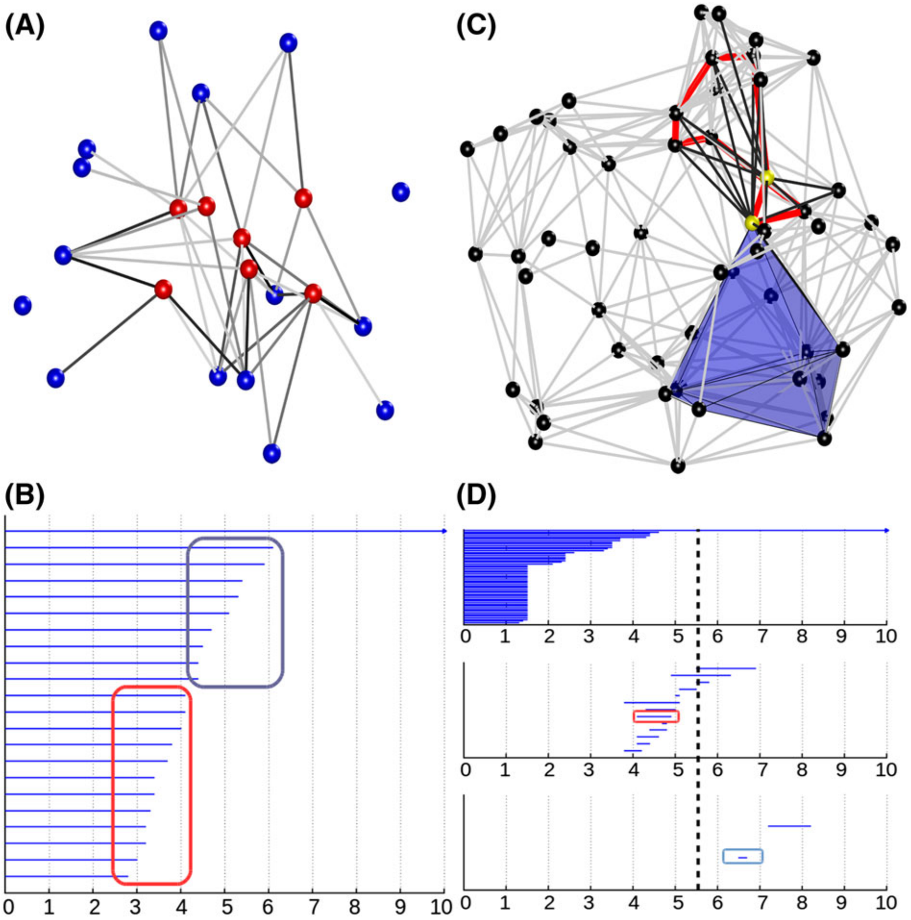


FIGURE 4 An illustration of element specific topological fingerprints (ESTF) indicating hydrophilic network (left) and hydrophobic network (right). The unit is Å for the barcode plots. A, Hydrophilic network showing the connectivity between nitrogen atoms of the protein (blue) and oxygen atoms of the ligand (red). B, The Betti-0 ESTFs of the aforementioned $N - O$ hydrophilic network. Betti-0 barcodes show not only the number and strength of of hydrogen bonds but also the hydrophilic environment. Specifically, bars in the left box can be directly interpreted as moderate or weak hydrogen bonds, while bars in the right box contributing the degree of hydrophilicity at the binding site. Black dash line shows the corresponding cutoff value of the connection network in A. C, Hydrophobic network constructed with carbon atoms near the binding site. D, The ESTFs of C highlighting protein-ligand hydrophobic contacts. Bars in the red box and the blue box correspond to the red loop and the blue cavity respectively in C. Black dash line shows the corresponding cutoff value of the connection network in C

kernel and the Lorentz kernel. The exponential kernel is defined as

$$\Phi^E(r;\eta_{ij},\kappa) = e^{-(r/\eta_{ij})^{\kappa}}, \tag{12}$$

and the Lorentz kernel is defined as

$$\Phi^L(r;\eta_{ij},\nu) = \frac{1}{1+\left(\frac{r}{\eta_{ij}}\right)^{\nu}}, \tag{13}$$

where k, τ, and ν are positive adjustable parameters that control the decay speed of the kernel allowing us to model interactions with difference strengths. Here, η_{ij} is the characteristic distance between the ith and the jth atoms and is usually set to be the sum of the van der Waals radii of the two atoms. The correlation between 2 atoms is then defined as

$$C_{ij} = \Phi(r_{ij}), \tag{14}$$

where r_{ij} is the Euclidean distance between the ith atom and the jth atom, and Φ is the kernel function. Note that the output of kernel functions lies in the $(0,1]$ interval. A correlation matrix is defined as

$$d(i,j) = 1 - C_{ij}. \tag{15}$$

The properties, $\Phi(0,\eta)=1$, $\Phi(r,\eta)\in(0,1], \forall r \geqslant 0$, $r_{ij}=r_{ji}$, and the strictly monotone decreasing property of the Φ assure the identity of indiscernible, nonnegativity, symmetry, and distance increases as pairwise interaction decays. Persistent homology computation is performed with VR complex built upon the afore-defined correlation matrix as an addition to the Euclidean space distance metric.

2.4.2 | TF/ESTF and feature construction

The TFs/ESTFs used in the machine-learning process are extracted from persistent homology computations with a variety of generalized metrics and different groups of atoms. First, the element type and atom center position of heavy atoms (nonhydrogen atoms) of both protein and ligand molecules are extracted. Hydrogen atoms are neglected because the procedure of completing protein structures by adding missing hydrogen atoms highly depends on the force field chosen, which will lead to force field depended effects. The point sets containing certain element types from the protein molecule and certain element types from the ligand molecule are grouped together. With this approach, the interactions between different element types are modeled separately and the parameters that distinguish between the interactions between different pairs of element types can be learned from the training set by machine-learning algorithms. The distance matrices with Euclidean distance and correlation matrix are constructed for each group of atoms. The features describing the TFs/ESTFs are then extracted from the outputs of persistent homology calculations and glued to form a feature vector for machine learning. The details of TF/ESTF construction and feature generation follow in the rest of this section.

2.4.3 | Groups of atoms

The types of elements considered for proteins are T_{P}={C, N, O, S} and those for ligands are T_{L} = {C, N, O, S, P, F, Cl, Br ,I}. We denote $P^c_{\mathrm{X-Y}}$ a set of atoms that consist of X type of atoms in protein and Y type of atoms in ligand, and the distance between any pair of atoms in these 2 groups is within a cutoff c:

$$P^c_{\mathrm{X-Y}} = \{a|a \in \mathrm{X}, \min_{b\in \mathrm{Y}} \mathrm{dis}(a,b) \leqslant c\} \cup \{b|b \in \mathrm{Y}\}, \tag{16}$$

where a and b denote atoms. As an example, $P^{12}_{\mathrm{C-O}}$ contains all O atoms in the ligand and all C atoms in the protein that are within the cutoff distance of 12Å from the ligand molecule. We denote the set of all heavy atoms in ligand together with all heavy atoms in protein that are within the cutoff distance c from the ligand molecule by P^c_{all}. Similarly, the set of all heavy atoms in protein that are within cutoff distance c from the ligand molecule by P^c_{pro}.

2.4.4 | Distance matrices

We define FRI-based correlation matrix and Euclidean (EUC) metric–based distance matrix in this section. The Lorentz kernel and exponential kernel defined in Equations 13 and 12, see Figure 5, are used due to their success in our previous studies of biomolecules with FRI theory. We use $A(i)$ to denote the affiliation of the atom with index i, which is either protein or ligand.

- $\mathrm{FRI}^{agst}_{\tau,\nu}$:

$$d(i,j) = \begin{cases} 1-(\frac{1}{1+(r_{ij}/\eta_{ij})^{\nu}}), & A(i) \neq A(j) \\ d_{\infty}, & A(i)=A(j) \end{cases}, \tag{17}$$

  where r_{ij} is the Euclidean distance between atoms with indices i and j and $\eta_{ij} = \tau * (r_i + r_j)$. Superscript $agst$ is the abbreviation of against and means that only the interaction between atoms in the protein and atoms in the ligand is taken into account since the binding between protein and ligand is studied and thus the distance between atoms from the same molecule is set to d_{∞}, which is a large positive number.

- $\mathrm{FRI}_{\tau,\nu}$:

$$d(i,j) = 1-\left(\frac{1}{1+(r_{ij}/\eta_{ij})^{\nu}}\right), \tag{18}$$

- EUC^{agst}:

$$d(i,j) = \begin{cases} r_{ij}, & A(i) \neq A(j) \\ d_{\infty}, & A(i)=A(j) \end{cases}, \tag{19}$$

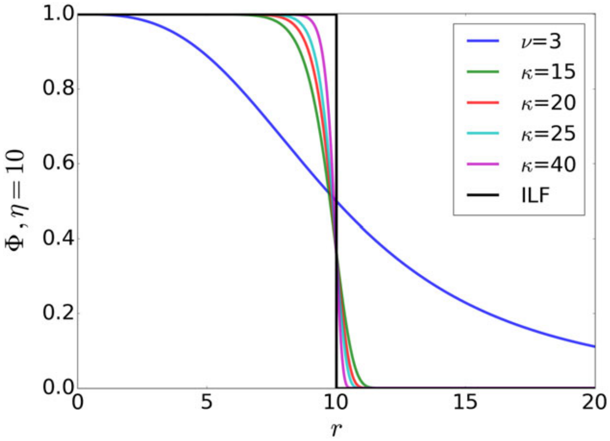


FIGURE 5 Illustration of flexibility-rigidity index correlation kernels at various powers. At a large κ or a very large ν, it essentially becomes an ideal low-pass filter and has a cutoff effect

TABLE 1 Feature extraction from TF/ESTF

Feature category	TF/ESTF	Features

Connectivity quantified with atomic interaction strengths.	ESTF($P^{12}_{\mathrm{C-C}}$, $\mathrm{FRI}^{agst}_{\Phi^L/\Phi^E}$, VR)
	$\cdots$
	ESTF($P^{12}_{\mathrm{S-I}}$, $\mathrm{FRI}^{agst}_{\Phi^L/\Phi^E}$, VR)
		Summation of all Betti-0 bar lengths.

	TF(P^{6}_{all}, $\mathrm{FRI}_{\Phi^L/\Phi^E}$, VR)
	TF(P^{6}_{pro}, $\mathrm{FRI}_{\Phi^L/\Phi^E}$, VR)
		Summation of length and birth of Betti-0, −1, and −2 TFs of protein, complex, and difference of the two.

Physical interactions grouped with intrinsic contact distance.	ESTF($P^{12}_{\mathrm{C-C}}$, EUC^{agst}, VR)
	$\cdots$
	ESTF($P^{12}_{\mathrm{S-I}}$, EUC^{agst}, VR)
		Counts of Betti-0 bars with “death” values falling into each interval: $[0, 2.5]$, $[2.5, 3]$, $[3, 3.5]$, $[3.5, 4.5]$, $[4.5, 6]$, $[6, 12]$.

Geometric features.	TF(P^{9}_{all}, Alpha)
	TF(P^{9}_{pro}, Alpha)
		Summation of Betti-1 and Betti-2 bar lengths with “birth” value falling into each intervals: $[0, 2]$, $[2, 3]$, $[3, 4]$, $[4, 5]$, $[5, 6]$, $[6, 9]$. The differences between the complex and protein are also taken into account.

Abbreviations: ESTF, element specific topological fingerprint; FRI, flexibility-rigidity index; TF, topological fingerprints; VR, Vietoris-Rips.

- EUC:

$$d(i,j) = r_{ij}. \tag{20}$$

The distance matrices with distance functions $\mathrm{FRI}^{agst}_{\tau,\nu}$ and EUC^{agst} mainly model connectivity and interaction between protein and ligand molecules and even higher order correlations between atoms. The Euclidean metric space is applied to detect geometric characteristics such as cavities and cycles.

We denote the output of persistent homology computation by TF(x, y, z), where x is the set of atoms; y is the distance matrix used; and z is the simplicial complex constructed. Notation ESTF(x, y, z) is used if x is element specific. The extraction of machine-learning feature vectors from TFs is summarized in Table 1.

The division of Betti-0 bars into bins is because that different categories of atomic interactions have their distinguish intrinsic distances such as 2.5 Å for ionic interactions and 3 Å for hydrogen bonds. The separation of Betti-1 and Betti-2 TFs helps grouping the geometric features of various scales. For FRI kernel, different pairs of parameters τ and ν are used to characterize interaction of different scales. In this work, the specific pairs of $[\tau, \nu]$ used are $[0.5, 3]$, $[1, 3]$, and $[2, 3]$.

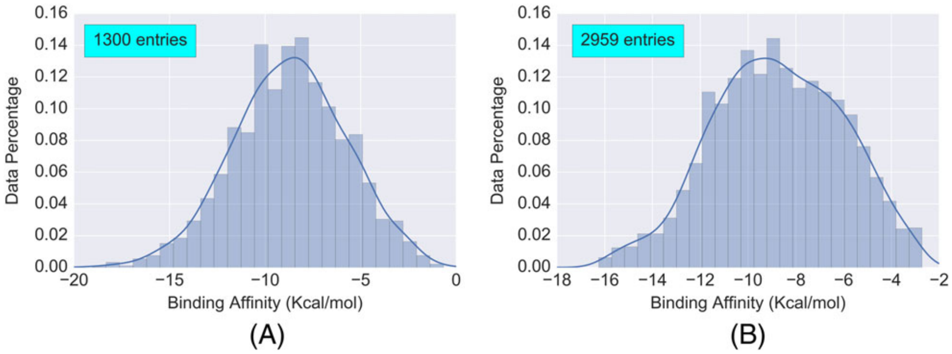


FIGURE 6 Distribution of binding affinity data in PDBBind v2007 refined set and PDBBind v2013 refined set

2.5 | Datasets

The PDBBind database provides a large collection of protein-ligand binding affinity data with atomic structure of protein-ligand complexes. The PDBBind v2007 refined set and PDBBind v2013 refined set are used to test the T-Bind binding affinity predictor. The distribution of the binding affinity data of 2 data sets is summarized in Figure 6. Core sets constructed by selecting three protein-ligand complexes with high, medium, and low binding affinity, respectively, for each protein are used to test the predicting power of the T-Bind predictor on a wide range of binding affinities. The core sets of both data sets consist of 195 complexes of 65 proteins but are not identical.

The protein-ligand binding affinity is reflected by the dissociation constant $K_d = [L][P]/[LP]$, where $[L]$, $[P]$, and $[LP]$ are the molar concentration of ligand, protein, and protein-ligand complex, respectively. The corresponding Gibbs free energy is $\Delta G = RT \ln K_d$ with R and T being the gas constant and temperature, respectively. At room temperature $T = 298.15K$, the energy can be computed as $\Delta G = -1.3633pK_d$ in the unit of kcal/mol. Here, pK_d is $-\log_{10} K_d$ with K_d in the unit of mol.

2.6 | Machine learning and validation

The gradient boosting trees regression of ensemble methods module (sklearn.ensemble.GradientBoostingRegressor) from Scikit-learn[48] is used to implement the proposed topological learning model. A uniform set of parameters used for all validation tasks is {n_estimator=20000, max_depth=8, min_samples_split=6, learning_rate=0.005, loss=ls, subsample=0.7, max_feature=sqrt}. Both predefined split of data into training and testing sets and fivefold cross validation of the data sets are used to validate the proposed method. To address the randomness coming from the machine-learning algorithm and the random splitting of data in fivefold cross validation, each experiment is repeated 50 times. The median performance and the best performance are reported together with the standard deviation of the performance across repeated experiments.

3 | RESULTS AND DISCUSSION

3.1 | General performance

To demonstrate the power of the proposed topological learning strategies for protein-ligand binding affinity prediction, we consider benchmark test sets from the PDBBind database.[49] This database provides a comprehensive collection of binding affinity data of biomolecular complexes available in Protein Data Bank (PDB)[50] that are obtained from experiments. We first consider the PDBBind v2007 refined set of 1300 protein-ligand complexes, which include the v2007 core set of 195 complexes.[51] Each protein-ligand complex has the 3-D structure of the biomolecule accompanied by the 3D structure of the ligand and the experimental binding affinity. Structures are directly imported to our topological learning predictor without any structure optimization, which tests the robustness of the predictor over data of relatively low quality. The v2007 core set has been used to evaluate the performance of more than 20 existing scoring functions and machine learning–based binding affinity predictor.[51-53] We use the refined set, excluding the core set, of 1105 complexes as the training set, while the core set of 195 complexes is used as the test set in our study.

Table 2 illustrates a comparison of Pearson correlation coefficients of the present method (T-Bind) and other 24 scoring functions given in the literature.[52] The proposed topological learning strategy outperforms all other methods, which

TABLE 2 The comparison of the proposed method named T-Bind with other scoring functions and machine learning–based method on the prediction of the PDBBind v2007 core set

PDBBind v2007 core set **Method**	R_P^a	**RMSE**^{b}
T-Bind	0.818	1.41
RF::VinaElem	0.803	1.42
RF::Vina	0.739	1.61
Cyscore	0.660	1.79
X-Score::HMScore	0.644	1.83
MLR::Vina	0.622	1.87
HYDE2.0::HbondsHydrophobic	0.620	1.89
PHOENIX	0.616	2.16
DrugScore^{CSD}	0.569	1.96
SYBYL::ChemScore	0.555	1.98
AutoDock Vina	0.554	1.99
DS::PLP1	0.545	2.00
GOLD::ASP	0.534	2.02
SYBYL::G-Score	0.492	2.08
DS::LUDI3	0.487	2.09
DS::LigScore2	0.464	2.12
GlideScore-XP	0.457	2.14
DS::PMF	0.445	2.14
GOLD::ChemScore	0.441	2.15
NHA-baseline	0.431	2.15
MWT-baseline	0.418	2.17
SYBYL::D-Score	0.392	2.19
IMP::RankScore	0.322	2.25
DS::Jain	0.316	2.24
GOLD::GoldScore	0.295	2.29
SYBYL::PMF-Score	0.268	2.29
SYBYL::F-Score	0.216	2.35

For the machine learning–based methods, T-Bind and RF::VinaElem, the training set is the PDBBind v2007 refined set minus v2007 core set. NHA-baseline is the number of heavy atoms and MWT-baseline is the weight of ligand, which serve as baseline methods. The data of all other methods are directly taken from Li et al.[54]

[a] Pearson correlation coefficient between computed values and experimental results.

[b] Root mean squared error in pKd units for the computed values.

represented the state of the art. The strong performance of the proposed topological learning strategy indicates the power of TFs/ESTFs for revealing protein-ligand binding mechanism, which is further explored in the rest of this paper. Similar comparison is made for PDBBind v2013 core set where the training set is PDBBind v2013 refined set minus the PDBBind v2013 core set. The comparison is summarized in Table 3.

A total of 6 tasks are performed over 3 datasets to further validate the T-Bind binding affinity predictor. Each test is performed 50 times with both the best performance and the median performance reported. The performance is summarized in Table 4. In the prediction of core sets, refined sets with their core sets stripped off are used as training sets. The five-fold cross-validation is performed with the random separation of the data set each time. Scatter plots of the predictions on different data sets are shown in Figure 7.

TABLE 3 The comparison of the proposed method named T-Bind with other scoring functions and machine learning–based method on the prediction of the PDBBind v2013 core set

PDBBind v2013 core set **Method**	R_P^a	**RMSE**^{b}
T-Bind	0.767	1.52
RF::VinaElem	0.752	1.49
X-Score^{HM}	0.614	1.78
ΔSAS	0.606	1.79
ChemScore@SYBYL	0.592	1.82
ChemPLP@GOLD	0.579	1.84
PLP1@DS	0.568	1.86
AutoDock Vina	0.564	1.86
G-Score@SYBYL	0.558	1.87
ASP@GOLD	0.556	1.88
ASE@MOE	0.544	1.89
ChemScore@GOLD	0.536	1.90
D-Score@SYBYL	0.526	1.92
Alpha-HB@MOE	0.511	1.94
LUDI3@DS	0.487	1.97
GoldScore@GOLD	0.483	1.97
Affinity-dG@MOE	0.482	1.98
NHA	0.478	1.98
MWT	0.473	1.99
LigScore2@DS	0.456	2.02
GlideScore-SP	0.452	2.03
Jain@DS	0.408	2.05
PMF@DS	0.364	2.11
GlideScore-XP	0.277	2.18
London-dG@MOE	0.242	2.19
PMF@SYBYL	0.221	2.20

For the machine learning–based methods, T-Bind and RF::VinaElem, the training set is the PDBBind v2013 refined set minus v2013 core set. NHA is the number of heavy atoms and MWT is the weight of ligand, which serve as baseline methods. The data of all other methods are directly taken from Li et al.[55]

[a] Pearson correlation coefficient between computed values and experimental results.

[b] Root mean squared error in pKd units for computed values.

3.2 | Feature analysis

It is interesting to analyze the predictive power of ESTFs. We consider different groups of features in the task of prediction of the PDBBind v2007 core set using models trained on the PDBBind v2007 refined set minus the core set. Typically, a more important ESTF has a higher predictive power. We first examine the performance of Betti-0 ESTFs of noncarbon heavy atoms that correspond to hydrogen bonds or hydrophilic interactions. This set of Betti-0 ESTFs delivers a Pearson correlation coefficient of 0.63 with a root-mean-square error (RMSE) of 2.51 kcal/mol, which is better than results of most other methods shown in Table 2. In contrast, Betti-0 ESTFs of carbon atoms only achieve a Pearson correlation coefficient of 0.59 with the RMSE of 2.62 kcal/mol. However, the mechanism of hydrophobic interactions differs much from that of hydrogen bonds, which are electrostatic in origin. Protein hydrophobic interactions are strengthened by the aggregation of nonpolar carbon atoms. Such aggregation is “uniquely” and effectively detected by carbon Betti-1 and

TABLE 4 The overall performance of the T-Bind binding affinity predictor on a variety of testing and cross-validation (CV) tasks

Dataset	Task	R_P^a	RMSE
v2007 refined	5-fold CV	0.749 (0.759) [0.005]	1.947 (1.914) [0.016]
v2007 refined	v2007 core	0.818 (0.823) [0.003]	1.918 (1.899) [0.010]
v2013 refined	5-fold CV	0.751 (0.757) [0.002]	1.798 (1.781) [0.007]
v2013 refined	v2013 core	0.767 (0.773) [0.003]	2.066 (2.049) [0.008]
v2015 refined	5-fold CV	0.770 (0.776) [0.002]	1.738 (1.721) [0.008]
v2015 refined	v2015 core	0.775 (0.782) [0.003]	2.032 (2.011) [0.008]

The average performance is reported with the best performance in the parenthesis and standard deviations in square brackets. R_P^a is the Pearson correlation coefficient and RMSE is the root mean squared error in the unit of Kcal/mol. Here, “v20xx core” stand for testing tasks.

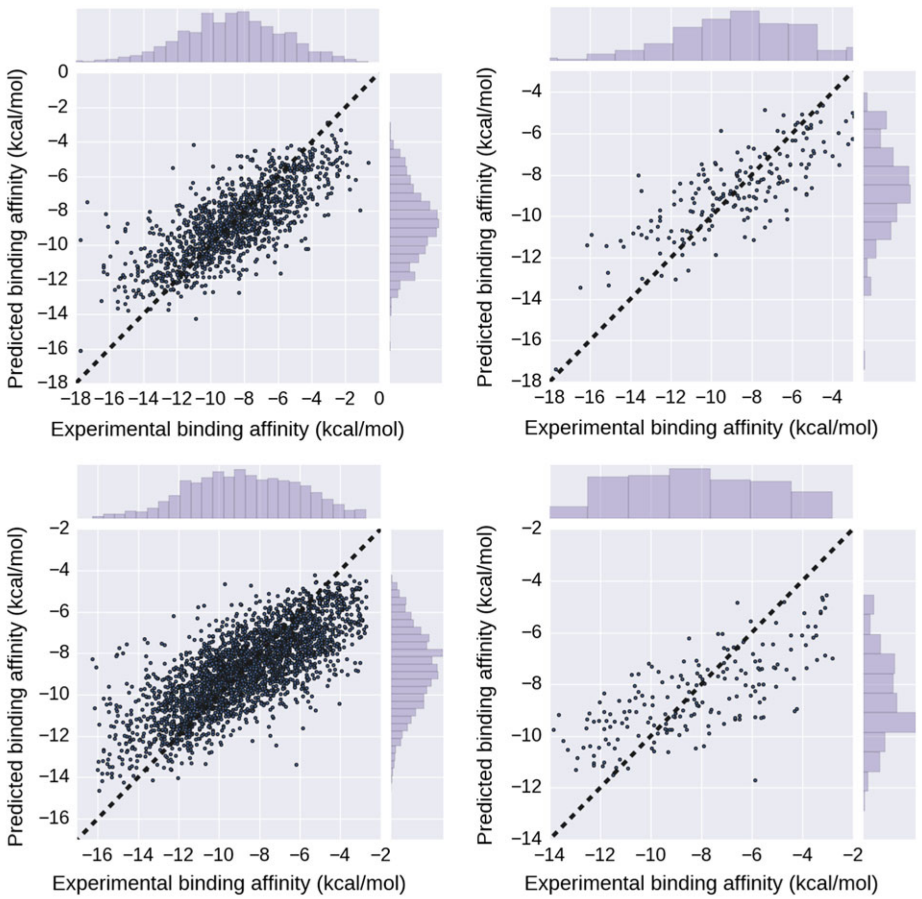


FIGURE 7 Correlation between T-Bind predictions and experimental data. Upper left, fivefold cross validation of the v2007 refined set. Upper right, prediction of the v2007 core set. Bottom left, fivefold cross validation of the v2013 refined set. Bottom right, prediction of the v2013 core set

Betti-2 ESTFs. Indeed, Betti-1 and Betti-2 ESTFs alone give rise to a Pearson correlation coefficient of 0.76 with an RMSE of 2.14 kcal/mol. The combination of all carbon ESTFs yields a respectable Pearson correlation coefficient of 0.79 with an RMSE of 2.04 kcal/mol. Therefore, hydrophobic interactions represented by carbon ESTFs have a higher predictive power than hydrophilic interactions represented by the ESTFs of noncarbon heavy atoms. The combination of Betti-0 ESTFs of all heavy atoms renders a remarkable Pearson correlation coefficient of 0.81 with RMSE of 1.97 kcal/mol. This result indicates that the characterization of hydrophobic effects is the most important to an accurate prediction. The additional consideration of hydrophilic interactions complements the features for hydrophobic interactions, and results in a marginal improvement on predictions, although the group of hydrophilic features delivers inferior results.

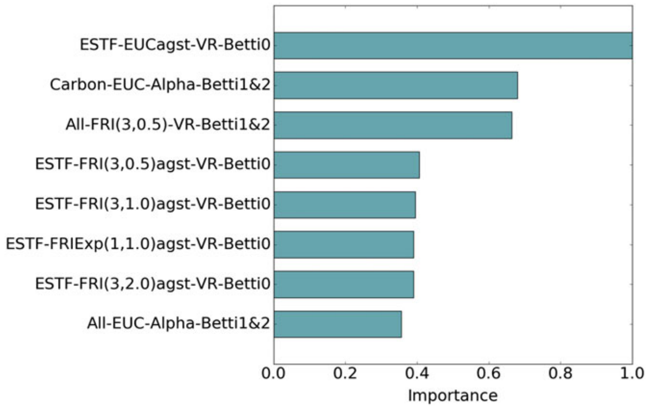


FIGURE 8 Feature importance for each group of features rescaled by dividing by the maximum importance score. Feature names are arranged as element types, distance function, simplicial complex, and Betti numbers. “ESTF" means that the 36 pairs of element types of protein and ligand are taken into account individually, and “All" means that all heavy atoms are considered as a single set. “FRI” is the FRI kernel with lorentz function, and “FRIExp” is the FRI kernel with exponential function. Among these feature groups, FRI-related groups are of the first category in Table 1, “ESTF-EUCagst-VR-Betti0” is of the second category, and “Carbon-EUC-Alpha-Betti1&2” and “All-EUC-Alpha-Betti1&2” are of the third category

The aforementioned results are obtained from ESTFs based on the standard Euclidean space distance filtration. Alternatively, a correlation function–based filtration can be used.[41] Specifically, the correlation function measures the distance between each pair of atoms with a specific scale so that certain distance is favored. One of correlation functions successfully used in protein flexibility analysis is Lorentz function $\left[1+\left(\frac{r_{ij}}{\eta_{ij}}\right)^{\nu}\right]^{-1}$, where r_{ij} is the distance between atom i and atom j, and $\eta_{ij} = \tau(r_i + r_j)$ is the scale.[45] Here, r_i and r_j are the van der Waals radii of atom i and atom j, respectively. Obviously, this filtration is regulated by a pair of parameters, ie, the correlation power ν and the correlation scale τ. It is found that Betti-0 ESTFs constructed from Lorentz function–based filtration are able to accurately predict binding affinities. Specifically, when $(\nu, \tau) = (3, 0.5), (3, 1), (3, 2), (2, 1)$, and $(5, 1)$, prediction Pearson correlation coefficients are $0.762, 0.758, 0.769, 0.777$, and 0.772, respectively. Obviously, other forms of correlation function can be used as well. For example, Betti-0 ESTFs constructed from exponential function $\exp\left[-\left(\frac{r_{ij}}{\eta_{ij}}\right)^{\kappa}\right]$–based filtration with $\kappa = \tau = 1$ delivers a Pearson correlation coefficient of 0.782. An interesting advantage of correlation function–based filtration is that multiscale effects in protein-ligand interactions can be captured by the use of multiple sets of ESTFs characterized at different scales. The combination of 3 sets of Betti-0 ESTFs $(\nu, \tau) = (3,0.5),(3,1)$ and $(3,2)$ leads to a high Pearson correlation coefficient of 0.784.

For the ensemble of trees method, the percentage of usage of a feature reflects its importance for the problem. The importance of each group of features used in this work is shown in Figure 8. All groups of features contribute to the model whilst contributions vary. Additional results for each group of features according to their element combinations are summarized in Table S1.

3.3 | Distance effect

We demonstrate that the present topological learning strategy is able to bring to light the effective lengths of various interactions. To this end, we consider exponential function–based filtration with a large κ value, ie, $\kappa = 15, 20, 25$ or 40, which results in an effective cutoff at the length scale of $\eta_{ij} = \tau(r_i + r_j)$.[56] Additional analysis of cutoff distance with alpha complex is discussed in Figure S1. Meanwhile, we consider 3 types of features, as shown in Figure 9. Type 1 shown in red includes all the Betti-0 features except those from protein-ligand C−C pairs. Type 2 shown in blue is obtained by restoring in type 1 Betti-0 features from protein-ligand C−C pairs. Type 3 shown in black is created by adding protein-ligand C−C Betti-1 and Betti-2 features to type 2. Interestingly, Figure 9 reveals molecular mechanism about protein-ligand binding. First, there are oscillatory peaks in Pearson correlation coefficients around $\tau = 1.6$, and 3.1, indicating the importance of complete inclusion of the first layer and/or the second layer of residues in predictive models. Additionally, it is interesting to note the importance of protein-ligand C−C interactions in nearest 2 layers of residues to binding. However, in the range of 3-5τ, the inclusion of protein-ligand C−C interactions appears to have little effect on binding predictions. Finally, at large length scales ($\tau > 5$), protein-ligand C−C Betti-1 and Betti-2 features give rise to a surprising catch-up contribution

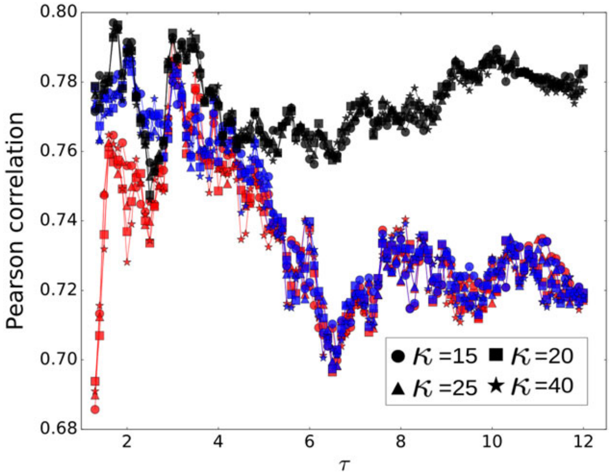


FIGURE 9 The Pearson correlation coefficient of type 1 (red), type 2 (blue), and type 3 (black) features with various cutoff distances achieved by varying length scale τ in predicting the PDBBind v2007 core set. All Betti-0 features are derived from VR complexes with the exponential correlation function with $\kappa = 15, 20, 25$, or 40. The Betti-1 and Betti-2 features are extracted from alpha complexes of protein-ligand C−C networks. The training/testing process is performed 50 times for each data point, and the average is taken

to binding predictions. This finding indicates that protein-ligand C−C network or protein-ligand hydrophobic effect is still significantly effective at more than 40Å away from the binding site, which has an important ramification to drug design and protein design.

4 | CONCLUSION

ESPH is introduced to simplify complex biomolecular geometry while retain crucial biological information. It provides a multicomponent persistent homology representation of biomolecules. ESPH is integrated with machine learning for the prediction of protein-ligand binding affinity utilizing 3-D structures. The outstanding performance of the proposed method suggests that persistent homology is able to deliver a comprehensive characterization of objects with complex geometries and can extract important features from complex biological units. The element specific topological fingerprint and the customized distance functions proposed in this work bridge conventional persistent homology and quantitative predictions for biomolecules. Since topological features are constructed with a physical flavor, the model can also be used to analyze the impacts of different physical interactions and characteristics on the target property. Specifically, we have shown that the impact of hydrophobic networks on the protein-ligand binding affinity may still be effective at a distance as long as 40 Å.

In spite of the thorough characterization of biomolecule structures, electrostatics, another crucial property of biomolecules, is not explicitly considered. The inclusion of electrostatics effects may deliver potential improvements. The performance of the proposed method on the prediction of protein-ligand binding affinity suggests its potential usage as a post processing function for selecting poses after protein-ligand docking analysis. It is interesting to formulate and benchmark the proposed method for the application to pose ranking for protein-ligand docking in the future.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grants DMS-1721024 and IIS-1302285 and MSU Center for Mathematical Molecular Biosciences Initiative.

ORCID

Zixuan Cang http://orcid.org/0000-0002-9951-5586
Guo-Wei Wei http://orcid.org/0000-0002-5781-2937

REFERENCES

1. Kollman PA, Massova I, Reyes C, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*. 2000;33:889-897.
2. Gilson MK, Given JA, Bush BL, McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J*. 1997;72(3):1047-1069.
3. Gilson MK, Zhou HX. Calculation of protein-ligand binding affinities. *Ann Rev Biophys Biomolecular Struct*. 2007;36:21-42.
4. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem*. 1995;38:2681-2691.
5. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV. Medusascore: an acurate force field-based scoring function for virtual drug screening. *J Chem Inf Model*. 2008;48:1656-1662.
6. Li H, Leung K-S, Wong M, Ballester PJ. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinform*. 2014;15(291).
7. Wang B, Zhao Z, Nguyen DD, Wei GW. Feature functional theory – binding predictor (FFT-BP) for the blind prediction of binding free energy. *Theor Chem Acc*. 2017;136:55.
8. Zheng Z, Merz KMJr. Ligand identification scoring algorithm (LISA). *J Chem Inf Model*. 2011;51:1296-1306.
9. Verkhivker G, Appelt K, Freer ST, Villafranca JE. Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus protease binding affinity. *Protein Eng*. 1995;8:677-691.
10. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*. 1997;11:425-445.
11. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure based binding affinity prediction. *J Comput Aided Mol Des*. 2002;16:11-26.
12. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom*. 2002;28:511-533.
13. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom*. 2005;33:249-274.
14. Kaczynski T, Mischaikow K, Mrozek M. *Computational Homology, vol. 157 of Applied Mathematical Sciences*. New York: Springer-Verlag; 2004.
15. Schlick T, Olson WK. Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA. *Science*. 1992;257(5073): 1110-1115.
16. Sumners DW. Knot theory and DNA. *Proc Symp Appl Math*. 1992;45:39-72.
17. Darcy IK, Vazquez M. Determining the topology of stable protein-DNA complexes. *Biochem Soc Trans*. 2013;41:601-605.
18. Heitsch C, Poznanovic S. Combinatorial insights into rna secondary structure. In: Jonoska N, Saito M, eds. *Discrete and Topological Models in Molecular Biology*, Vol. Chapter 7, Springer Berlin Heidelberg; 2014:145-166.
19. Demerdash ONA, Daily MD, Mitchell JC. Structure-based predictive models for allosteric hot spots. *PLOS Comput Biol*. 2009;5:e1000531.
20. DasGupta B, Liang J. *Models and Algorithms for Biomolecules and Molecular Networks*: John Wiley & Sons: Hoboken, New Jersey; 2016.
21. Shi X, Koehl P. Geometry and topology for modeling biomolecular surfaces. *Far East J Appl Math*. 2011;50:1-34.
22. Bendich P, Harer J. Persistent intersection homology. *Found Comput Math (FOCM)*. 2011;11(3):305-336.
23. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. *Discrete Comput Geom*. 2007;37(1):103-120.
24. Cohen-Steiner D, Edelsbrunner H, Harer J. Extending persistence using poincaré and lefschetz duality. *Found Comput Math*. 2009;9(1):79-103.
25. Cohen-Steiner D, Edelsbrunner H, Harer J, Morozov D. Persistent homology for kernels, images, and cokernels. In: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 09, New York, New York, USA; 2009:1011-1020.
26. Chazal F, Cohen-Steiner D, Glisse M, Guibas LJ, Oudot S. Proximity of persistence modules and their diagrams. In: Proc. 25th ACM Sympos. on Comput. Geom, Aarhus, Denmark; 2009:237-246.
27. Chazal F, Guibas LJ, Oudot SY, Skraba P. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)* 2013;60(6):41.
28. Carlsson G, Zomorodian A. The theory of multidimensional persistence. *Discrete Comput Geom*. 2009;42(1):71-93.
29. Carlsson G, de Silva V, Morozov D. Zigzag persistent homology and real-valued functions. In: Proc. 25th Annu. ACM Sympos. Comput. Geom, Aarhus, Denmark; 2009:247-256.
30. de Silva V, Morozov D, Vejdemo-Johansson M. Persistent cohomology and circular coordinates. *Discrete Comput Geom*. 2011;45:737-759.
31. Carlsson G, De Silva V. Zigzag persistence. *Found Comput Math*. 2010;10(4):367-405.
32. Oudot SY, Sheehy DR. Zigzag zoology: Rips zigzags for homology inference. *Foundations of Computational Mathematics* 2015;15(5):1151-1186.
33. Dey TK, Fan F, Wang Y. Computing topological persistence for simplicial maps. In: Proc. 30th Annu. Sympos. Comput. Geom. (SoCG), Kyoto, Japan; 2014:345-354.
34. Mischaikow K, Nanda V. Morse theory for filtrations and efficient computation of persistent homology. *Discrete Comput Geom*. 2013;50(2):330-353.
35. Adams H, Tausz A, Vejdemo-Johansson M. JavaPlex: A research software package for persistent (co) homology. *International Congress on Mathematical Software*. Springer; 2014:129-136.

36. Nanda V. Perseus: the persistent homology software. 2012. Software available at http://www.sas.upenn.edu/vnanda/perseus

37. Bauer U, Kerber M, Reininghaus J. Distributed computation of persistent homology. In: Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX), Portland, Oregon, USA; 2014: 31-38.

38. Xia KL, Wei GW. Persistent topology for cryo-EM data analysis. *Int J Numer Method Biomed Eng*. 2015;31:e02719.

39. Morozov D. Dionysus library for computing persistent homology; 2012.

40. Yao Y, Sun J, Huang XH, et al. Topological methods for exploring low-density states in biomolecular folding pathways. *J Chem Phys*. 2009;130:144115.

41. Xia KL, Wei GW. Persistent homology analysis of protein structure, flexibility and folding. *Int J Numer Method Biomed Eng*. 2014;30:814-844.

42. Mate G, Hofmann A, Wenzel N, Heermann DW. A topological similarity measure for proteins. *Biochim Biophys Acta - Biomembr*. 2014;1838:1180-1190.

43. Cang Z, Mu L, Wu K, Opron K, Xia K, Wei G-W. A topological approach to protein classification. *Mol Based Math Biol*. 2015;3:140-162.

44. Ghrist R. Barcodes: The persistent topology of data. *Bull Amer Math Soc*. 2008;45:61-75.

45. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models—Flexibility and rigidity. *J Chem Phys*. 2013;139:194-109.

46. Opron K, Xia KL, Wei GW. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *J Chem Phys*. 2014;140:234105.

47. Xia KL, Feng X, Tong YY, Wei GW. Persistent homology for the quantitative prediction of fullerene stability. *J Comput Chem*. 2015;36:408-422.

48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Machine Learn Res*. 2011;12:2825-2830.

49. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*. 2015;31(3):405-412.

50. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):35-242.

51. Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*. 2009;49:1079-1093.

52. Li W, Cowley A, Uludag M, et al. The embl-ebi bioinformatics web and programmatic tools framework. *Nucleic Acids Res*. 2015; 34:gkv279.

53. Li G-B, Yang L-L, Wang W-J, Li L-L, Yang S-Y. ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J Chem Inf Model*. 2013;53(3):592-600.

54. Li H, Leung K-S, Wong M-H, Ballester PJ. Improving autodock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular Inform*. 2015;34(2-3):115-126.

55. Li H, Leung K-S, Wong M-H, Ballester PJ. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*. 2015;20:10947-10962.

56. Xia KL, Opron K, Wei GW. Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). *J Chem Phys*. 2015;143:204106.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Cang Z, Wei G-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Meth Biomed Engng*. 2018;34:e2914. https://doi.org/10.1002/cnm.2914