PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening

Zixuan Cang[1], Lin Mu[2], Guo-Wei Wei[1,3,4]*

1 Department of Mathematics, Michigan State University, East Lansing, Michigan, United States of America, **2** Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **3** Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, United States of America, **4** Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, United States of America

* wei@math.msu.edu

Citation: Cang Z, Mu L, Wei G-W (2018) Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Comput Biol 14(1): e1005929. https://doi.org/10.1371/journal.pcbi.1005929

Editor: Jian Peng, University of Illinois at Urbana-Champaign, UNITED STATES

Received: September 1, 2017

Accepted: December 15, 2017

Published: January 8, 2018

Data Availability Statement: The PDBBind datasets can be found at "http://www.pdbbind.org.cn/". The S1322 set which is a subset of PDBBind dataset can be found in the supplementary material of the paper "Wang B, Zhao Z, Nguyen D D, et al. Feature functional theory–binding predictor (FFT–BP) for the blind prediction of binding free energies [J]. Theoretical Chemistry Accounts, 2017, 136(4): 55.". The DUD dataset with recalculated charge can be found at "http://dud.docking.org/inhibox.html" with the corresponding publication "Armstrong M S, Morris G M, Finn P W, et al. ElectroShape: fast

Abstract

This work introduces a number of algebraic topology approaches, including multi-component persistent homology, multi-level persistent homology, and electrostatic persistence for the representation, characterization, and description of small molecules and biomolecular complexes. In contrast to the conventional persistent homology, multi-component persistent homology retains critical chemical and biological information during the topological simplification of biomolecular geometric complexity. Multi-level persistent homology enables a tailored topological description of inter- and/or intra-molecular interactions of interest. Electrostatic persistence incorporates partial charge information into topological invariants. These topological methods are paired with Wasserstein distance to characterize similarities between molecules and are further integrated with a variety of machine learning algorithms, including k-nearest neighbors, ensemble of trees, and deep convolutional neural networks, to manifest their descriptive and predictive powers for protein-ligand binding analysis and virtual screening of small molecules. Extensive numerical experiments involving 4,414 protein-ligand complexes from the PDBBind database and 128,374 ligand-target and decoy-target pairs in the DUD database are performed to test respectively the scoring power and the discriminatory power of the proposed topological learning strategies. It is demonstrated that the present topological learning outperforms other existing methods in protein-ligand binding affinity prediction and ligand-decoy discrimination.

Author summary

Conventional persistent homology neglects chemical and biological information during the topological abstraction and thus has limited representational power for complex chemical and biological systems. In terms of methodological development, we introduce advanced persistent homology approaches for the characterization of small molecular

molecular similarity calculations incorporating shape, chirality and electrostatics[J]. Journal of computer-aided molecular design, 2010, 24(9): 789-801." The code for feature generation is included as supporting material (S1 Code).

Funding: Funds were received from National Science Foundation IIS-1302285 (https://www.nsf.gov/div/index.jsp?div=IIS) to GW, National Science Foundation DMS-1721024 (https://www.nsf.gov/div/index.jsp?div=DMS) to GW, and Michigan State University (https://vprgs.msu.edu/) to GW. The MSU funding is a combination of general supports (for example, teaching duty is reduced for mentoring undergrad students) from the school and therefore, we can not provide a specific grant number. The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

structures which can capture subtle structural difference. We also introduce electrostatic persistent homology to embed physics in topological invariants. These approaches encipher physics, chemistry and biology, such as hydrogen bonds, electrostatics, van der Waals interactions, hydrophobicity and hydrophilicity, into topological fingerprints which, although cannot literally recast into physical interpretations, are ideally suitable for machine learning, particularly deep learning, rendering topological learning algorithms. In terms of applications, we construct a structure-based virtual screening model which outperforms other existing methods. This competitive model on the DUD database is derived by assessing the performance of a comprehensive collection of topological approaches proposed in this work and introduced in our earlier work, on the PDBBind database. The topological features constructed in this work can readily be applied to other biomolecular problems where the characterization of proteins or small molecules is needed.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Arguably, machine learning has become one of the most important developments in data science and artificial intelligence. With its ability to extract features of various levels hierarchically, deep convolutional neural networks (CNNs) have made breakthroughs in image processing, video, audio, and computer vision [1, 2], whereas recurrent neural networks have found success in analyzing sequential data, such as text and speech [3–6]. Deep learning algorithms are able to automatically extract high-level features and discover intricate patterns in large data sets. In general, one of the major advantages of machine learning algorithms is their ability to deal with large and diverse data sets and uncover complicated relationships.

Recently, machine learning has become an indispensable tool in biomolecular data analysis and structural bioinformatics. Almost every computational problem in molecular biophysics and biology, such as the predictions of solvation free energy, solubility, partition coefficient, protein-ligand binding affinities, mutation induced protein stability change, molecular multipolar electrostatics, virtual screening, etc., has machine learning based approaches that are either parallel or complementary to their physics based counterparts. The success of deep learning has fueled the rapid growth in several areas of biological science [3, 5, 6], including bioactivity of small-molecule drugs [7–10] and genetics [11, 12], where large data sets are available.

A key component of a learning machine based on biomolecular structures is featurization, that is translating the 3D structures of biomolecules to features. While the degrees of freedom of the original biomolecular structures are large and vary among different molecules, it is almost inevitable that information loss happens with dimension reduction during featurization. Besides the choice of learning models, the performance of a predictor heavily depends on how the features are extracted. Although deep learning has been known to be powerful for the automatic extraction of features from original inputs such as images, deep learning based models directly taking biomolecules as inputs are not as competitive as the state-of-art machine

learning models with carefully designed features, due to the intrinsic complexity of biomolecules [13].

Biomolecules can be characterized by geometric features, electrostatic features, high-level (residue and global level) features, and amino-acid sequence features based on physical, chemical, and biological understandings [14]. Geometric features, such as coordinates, distances, angles, surface areas [15–17] and curvatures [18–21], are important descriptors of biomolecules [22–24]. However, geometric features often involve too much structural detail and are frequently computationally intractable for large biomolecular data sets. Electrostatic features include atomic partial charges, Coulomb potentials, atomic electrostatic solvation energies, and polarizable multipolar electrostatics [25]. These descriptors become essential for highly charged biomolecular systems, such as nucleic acid polymers and some protein-ligand complexes. High-level features refer to pKa values of ionizable groups and neighborhood amino acid compositions, such as the involvement of hydrophobic, polar, positively charged, negatively charged, and special case residues. Sequence features consist of secondary structures, position-specific scoring matrix (PSSM), and co-evolution information. Sequence features and annotations provide a rich resource for bioinformatics analysis of biomolecular systems. Topology offers a new unconventional representation of biomolecules. Topology can describe biomolecules in a variety of ways [26]. Some of the most powerful topological features are obtained from multi-component persistent homology or element specific persistent homology (ESPH) [14, 27]. Recently, we carried out a comprehensive comparison of the performance of geometric features, electrostatic features, high-level features, sequence features and topological features, for the prediction of mutation induced protein folding free energy changes of four mutation data sets [14]. Surprisingly, topological features outperform all the other features [14].

Unlike geometry, topology is well known for its power of simplification to geometric complexity [28–35]. The global description generated by classical topology is based on the concept of neighborhood and connectedness. If a space can be continuously deformed to another, they are considered to have the same topological features. In this sense, topology can not distinguish between a folded protein and its unfolded form if only covalent bonds are considered. Such property prevents the use of classical topology for the characterization of biomolecular structures. Instead of using topology to describe a single configuration of connectivity, persistent homology scans over a sequence of configurations induced by a filtration parameter and renders a sequence of topological invariants, which partially captures part of geometric features. Persistent homology has been applied to biomolecular systems in our earlier works [26].

In mathematics, persistent homology is a relatively new branch of algebraic topology [29, 36]. When dealing with proteins and small molecules, it is conventional to consider atoms as point clouds. For a given point cloud data set, one type of persistent homology turns each point into a sphere with their radii systematically increasing. The corresponding topological invariants and their persistence over the varying radius values can be computed. Therefore, this method embeds multiscale geometric information in topological invariants to achieve an interplay between geometry and topology. Consequently, persistent homology captures topological structures continuously over a range of spatial scales. It is called persistent homology because at each given radius, topological invariants, i.e., Betti numbers, are practically calculated by means of homology groups. In the past decade, much theoretical formulation [37–46] and many computational algorithms [47–52] have been developed. One-dimensional (1D) topological invariants generated from persistent homology is often visualized by persistence barcodes [53, 54] and persistence diagrams [55]. In recent years, multidimensional persistence has attracted much attention [43, 56] in hope that it can better characterize the data shape when there are multiple measurements of interest.

Persistent homology has been applied to various fields, including image/signal analysis [57–62], chaotic dynamics verification [63, 64], sensor networks [65], complex networks [66, 67], data analysis [68–72], shape recognition [73–75], and computational biology [76–79]. Compared with traditional computational topology [80–82] and/or computational homology, persistent homology inherently adds an additional dimension, i.e., the filtration parameter. The filtration parameter can be used to embed important geometric or quantitative information into topological invariants. As such, the importance of retaining geometric information in topological analysis has been recognized [83], and persistent homology has been advocated as a new approach for handling big and high dimensional data sets [54, 68, 84–86]. Recently, we have introduced persistent homology for mathematical modeling and/or prediction of nanoparticles, protein unfolding, and other aspects of biomolecules [26, 87]. We proposed the molecular topological fingerprint (TF) to reveal *topology-function relationships* in protein folding and protein flexibility [26]. We established some of the first quantitative topological analyses in our persistent homology based predictions of the curvature energy of fullerene isomers [87, 88]. We have also shown correlation between persistence barcodes and energies computed with physical models during molecular dynamics experiments [26]. Moreover, we have introduced the first differential geometry based persistent homology that utilizes partial differential equations (PDEs) in filtration [88]. Most recently, we have developed a topological representation to address additional measurements of interest, by stacking the persistent homology outputs from a sequence of frames in molecular dynamics or a sequence of different resolutions [89, 90]. We have also introduced one of the first uses of topological fingerprints for resolving ill-posed inverse problems in cryo-EM structure determination [91]. In 2015, we constructed one of the first integrations of topology and machine-learning and applied it to protein classification involving tens of thousands of proteins and hundreds of tasks [92]. We also developed persistent-homology based software for the automatic detection of protein cavities and binding pockets [93].

Despite much success, it was found that persistent homology has a limited characterization power for proteins and protein complexes, when applied directly to biomolecules [92]. Essentially, biomolecules are not only complex in their geometric constitution, but also intricate in biological constitution. In fact, the biological constitution is essential to biomolecular structure and function. Persistent homology that is designed to reduce the geometric complexity of a biomolecule neglects biological information. To overcome this difficulty, we have introduced multi-component persistent homology or element specific persistent homology (ESPH) to recognize the chemical constitution during the topological simplification of biomolecular geometric complexity [14, 27, 94]. In ESPH, the atoms of a specific set of element types in a biomolecule are selected so that specific chemical information, such as hydrophobicity or hydrophilicity, is emphasized in each selection. Our ESPH is not only able to outperform other geometric and electrostatic representations in large and diverse data sets, but is also able to shed light on the molecular mechanism of protein-ligand binding, such as the relative importance of hydrogen bond, hydrophilicity and hydrophobicity at various spatial ranges [27].

The objective of the present work is to further explore the representability and reduction power of multi-component persistent homology for biomolecules and small molecules. To this end, we take a combinatorial approach to scan a variety of element combinations and examine the characterization power of these components. Additionally, we also propose a multi-level persistence to study the topological properties of non-covalent bond interactions. This approach enables us to devise persistent homology to describe the interactions of interest between atoms that are connected by weak non-covalent bonds and delivers richer representation especially for small molecules. Moreover, realizing that electrostatics are of paramount importance in biomolecules and to enhance the power of our topological representation, we

introduce electrostatic persistence, which embeds charge information in topological invariants, as a new class of features in multi-component persistent homology. The aforementioned approaches can be realized via the modification of the distance matrix with a more abstract setting, for example, Vietoris-Rips complex. The complexity reduction is guaranteed in the 1D topological representation of 3D biomolecular structures. Obviously, the multi-component persistent homology representation of biomolecule leads to a higher machine learning dimensionality compared to the original single component persistent homology for a biomolecule. Therefore, it is subject to overfitting or overlearning problem in machine learning theory. Fortunately, gradient boosting trees (GBT) method is relatively insensitive to redundant high dimensional topological features [14]. Finally, since the components can be arranged as a new dimension ordered by their feature importance, multi-component persistent homology barcodes are naturally a two-dimensional (2D) representation of biomolecules. Such a 2D representation can be easily used as image-like input data in a deep CNN architecture, with different topological dimensions, i.e., 0, 1, and 2, being treated as channels. Such a topological deep learning approach addresses the nonlinear interactions among important element combinations while keeping the information from less important ones. Barcode space metrics, such as bottleneck distance and more generally, Wasserstein distance [95, 96], offer a direct description of similarity between molecules and can be readily used with nearest neighbor regression or kernel based methods. The performance of Wasserstein distance for protein-ligand binding affinity predictions is examined in this work.

After assessing the new method's ability to represent small molecules and protein-compound complexes, the derived model is used for virtual screening. Virtual screening computationally screens a collection of small molecules to identify those who can potentially bind to the protein target. There are mainly two types of virtual screening which are ligand-based and structure-based. Ligand-based approaches depend on a measurement of similarity among small molecules using either 2D or 3D structural information of small molecules. Structure-based approaches attempt to dock the small molecule candidate to the protein target and determine if the candidate is a potential ligand based on the top docking poses. The performance of structure-based virtual screening methods heavily depends on the quality of the docking method and the accuracy of the post-docking scoring method. Our effort focuses on the development of a topology based method for the latter part. It has been shown that using machine learning or deep learning based methods to rescore the docking poses can significantly boost the performance [97, 98]. For the models such as ensemble of trees and classical neural networks, carefully constructed features are needed. For example, a neural network based method NNScore uses a collection of derived features such as the count of hydrogen bonds and electrostatics of close contacts to describe the protein-compound complex [97]. Another class of deep learning based methods feed lower level features to deep neural networks and relies on the neural networks to automatically extract higher-level features. For example, DeepVS first computes features on each atom involved in the docking interface and feed this information to a deep neural network starting with convolution layers to hierarchically extract higher-level features [98].

The rest of this manuscript is organized as follows. *Section Methods* is devoted to introducing methods and algorithms. We present multi-component persistent homology, multi-level interactive persistent homology, vectorized persistent homology representation and electrostatic persistence. These formulations are crucial for the representability of persistent homology for biomolecules. Machine learning algorithms associated with the present topological data analysis are briefly discussed. Results are presented in *Section Results*. We first consider the characterization of small molecules. More precisely, the cross-validation of protein-ligand binding affinities prediction via solely ligand topological fingerprints is studied. We illustrate

the excellent representability of our multi-component persistent homology by a comparison with a method using physics based descriptors. Additionally, we investigate the representational power of the proposed topological method on a few benchmark protein-ligand binding affinity data sets, namely, PDBBind v2007, PDBBind v2013, PDBBind v2015 and PDBBind v2016 [99]. These data sets contain thousands of protein-ligand complexes and have been extensively studied in the literature. Results indicate that multi-component persistent homology offers one of most powerful representations of protein-ligand binding systems. The aforementioned study of the characterization of small molecules and protein-ligand complexes leads to an optimal selection of features and models to be used for virtual screening. Finally, we consider the directory of useful decoys (DUD) database to examine the representability of our multi-component persistent homology for virtual screening to distinguish actives from non-actives. The DUD data set used in this work has a total of 128,374 ligand-target and decoy-target pairs containing 3961 active ligand-target pairs, and involves 40 protein targets from six families. A large number of state-of-the-art virtual screening methods have been applied to this data set. We demonstrate that the present multi-component persistent homology outperforms other methods with reported results on this benchmark. This paper ends with a conclusion.

Results

Rational drug design and discovery have rapidly evolved into some of the most important and exciting research fields in medicine and biology. These approaches potentially have a profound impact on human health. The ultimate goal is to determine and predict whether a given drug candidate will bind to a target so as to activate or inhibit its function, which results in a therapeutic benefit to the patient. Virtual screening is an important process in rational drug design and discovery which aims to identify actives of a given target from a library of small molecules. There are mainly two types of screening techniques, ligand-based and structure-based. Ligand-based approaches depend on the similarity among small molecule candidates. Structure-based approaches try to dock a candidate molecule to the target protein and judge the candidate with the modeled binding affinity based on docking poses. Various molecular docking software packages have been developed for these purposes. Molecular docking involves both pose generation and binding affinity scoring. Currently, pose generation is quite robust while scoring power is still limited. Therefore, knowledge-based rescoring methods using machine learning or deep learning approaches can improve scoring accuracy [97, 98, 100]. We also apply our topological learning method as a rescoring machine to rerank the candidates based on docking poses generated by docking software.

This section explores the representational power of the proposed persistent homology methods for the prediction of protein-ligand binding affinities and the discrimination of actives and non-actives for protein targets. To this end, we use the present method to investigate three types of problems. First, we develop topological learning models for ligand based protein-ligand binding affinity predictions. This problem is designed to examine the representability of the proposed topological methods for small molecules. Then, we develop topological learning models for protein-ligand complex based binding affinity prediction. This problem enables us to understand the capability of the proposed topological learning methods for dealing with protein-ligand complexes. Finally, we examine the structure-based classification of active ligands and decoys which are highly possible to be non-actives, i.e., structure-based virtual screening (VS). The optimal selection of features and methods are determined by studying the first two applications and this finding leads to the main application studied in this work, the topological structure-based virtual screening. Computational algorithms used in this study are illustrated in Fig 1.

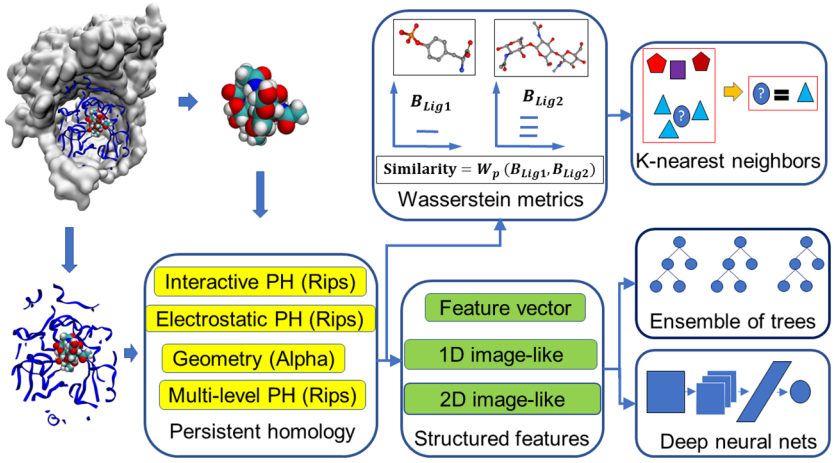

Fig 1. An illustration of the topology based machine learning algorithms used in scoring and virtual screening.

https://doi.org/10.1371/journal.pcbi.1005929.g001

Ligand based protein-ligand binding affinity prediction

In this section, we address the representation of small molecules by element specific persistent homology, especially the proposed multi-level persistent homology designed for small molecules.

Data set. To assess the representational ability of the present persistent homology algorithms on small molecules, we use a high quality data set of 1322 protein-ligand complexes with binding affinity data involving 7 protein clusters introduced earlier (denoted as S1322) [101]. It is a subset of the PDBBind v2015 refined set and its detail is given in the Supplementary material 1 of Ref. [101]. We consider a ligand based approach to predict the binding affinities of protein-ligand complexes in various protein clusters. As such, only the ligand information is used in our topological analysis. The ligand structures are taken from PDBBind database without modification. Numbers of ligands in protein clusters range from 94 to 333.

Models and performance. Two models, i.e., TopBP-KNN(Ligand) and TopBP-ML (Ligand), are constructed. TopBP-KNN(Ligand) is used to directly assess the representation power of persistent homology for small molecules and TopBP-ML(Ligand) is the final practical model. The results are shown in Table 1. All the gradient boosting trees models take the setup described in *Section Methods/Machine learning algorithms/Gradient boosting trees*.

In TopBP-ML(Ligand), we process the geometry, the shape, and the covalent bond information of the small molecules using alpha complex, and the non-covalent intramolecular interactions using multi-level persistent homology with Rips complex. The features used are

Table 1. Pearson correlation coefficients (RMSE in kcal/mol) of ligand based topological model on the S1322 dataset.

Methods	CL 1 (333)	CL 2 (264)	CL 3 (219)	CL 4 (156)	CL 5 (134)	CL 6 (122)	CL 7 (94)	Average
TopBP-KNN(Ligand)	0.698(1.66)	0.817(1.28)	0.620(1.68)	0.645(1.41)	0.756(1.68)	0.658(1.68)	0.739(1.31)	0.705(1.49)
TopBP-ML(Ligand) (5-fold)	0.713(1.60)	0.843(1.15)	0.693(1.51)	0.670(1.35)	0.831(1.34)	0.698(1.56)	0.737(1.26)	0.741(1.40)
FFT-BP (5-fold) [101]	(1.93)	(1.32)	(2.01)	(1.61)	(2.02)	(2.06)	(1.71)	(1.81)

The numbers in the first row show the number of entries in each protein cluster. The performance is reported as Pearson correlation coefficient (root mean squared error in kcal/mol). The median performance of 20 random 5-fold cross validation results is reported for TopBP-ML(Ligand). The results reported for TopBP-KNN (Ligand) are obtained by leave-one-out validation within each protein cluster with $k = 3$ for the KNN model.

https://doi.org/10.1371/journal.pcbi.1005929.t001

A-B012-E-S-GBT and R-B012-M1-S-GBT as described in *Section Discussion/Ligand based protein-ligand binding affinity prediction*. Gradient boosting trees method is used.

In TopBP-KNN(Ligand), we represent the small molecules with a collection of barcodes from element specific persistent homology calculations. Wasserstein distance with $p = 2$ is applied to measure similarities between two barcodes. The similarity between each pair of small molecules is then measured by taking the average of the Wasserstein distances between all considered barcodes. K-nearest-neighbor (KNN) regression is then applied to the measured similarity. In detail, the 6 barcodes considered are, R-B0-E-KNN, R-B1-E-KNN, R-B2-E-KNN, R-B0-M1-KNN, R-B1-M1-KNN, and R-B2-M1-KNN as described in *Section Discussion/Ligand based protein-ligand binding affinity prediction*. Leave-one-out validation within each protein cluster with $k = 3$ is used for this model.

In Table 1, FFT-BP 5-fold cross validation results were obtained based on multiple additive regression trees and a set of physical descriptors, including geometry, charge, electrostatic interactions, and van der Waals interactions for S1322 set [101]. Since multiple additive regression trees is also an implementation of the GBT used in the present work, it is appropriate to compare the FFT-BP results with the GBT results in this work to assess representation power of topological features. It is interesting to note that judging by RMSE, both sets of current topological descriptors have more predictive power than the physical descriptors built on protein-ligand complexes constructed in our earlier work [101]. These physical descriptors were constructed from sophisticated surface areas, molecular volumes, van der Waals interactions, charges computed by quantum mechanics, and Poisson-Boltzmann theory based electrostatics [101]. The success of topological descriptors implies the existence of an alternative and potentially more powerful description of the complex biomolecular world.

Complex based protein-ligand binding affinity prediction

In this section, we develop topological representations of protein-ligand complexes.

Data sets. The PDBBind database provides a comprehensive collection of structures of protein-ligand complexes and their binding affinity data [99, 102]. The original experimental data in Protein Data Bank (PDB) [103] are selected to PDBBind database based on certain quality requirements and are curated for applications. As shown in Table 2, this database is expanding on a yearly basis. It has become a common resource for benchmarking computational methods and algorithms for protein-ligand binding analysis and drug design. Popular data sets include version 2007 (v2007), v2013, and v2015. Among them, v2013 core set and v2015 core set are identical. A large number of scoring functions has been tested on these data sets. The latest version, v2016, has an enlarged core set, which contains 290 protein-ligand complexes from 58 protein families. Therefore, this test set should be relatively easier than v2015 core set, whose 195 complexes involve 65 protein families. The core sets are constructed by choosing 3 samples with median, maximum, and minimum binding affinity from each

Table 2. Description of the PDBBind datasets.

Version	Refined set	Training set	Core set (test set)	Protein families
v2007	1300	1105	195	65
v2013	2959	2764	195	65
v2015	3706	3511	195	65
v2016	4057	3767	290	58

Number of complexes or number of protein families in PDBBind data sets used in the present binding affinity prediction. Here training sets are set to the corresponding refined sets, excluding the complexes in the corresponding test sets (i.e., core sets). Protein families refer to those in the corresponding core sets.

https://doi.org/10.1371/journal.pcbi.1005929.t002

protein family for v2007, v2013, and v2015 sets. The core set for v2016 was constructed similarly but with 5 samples from each protein family.

Model and performance. Two models TopBP-ML(Complex) and TopBP-DL(Complex) are introduced. The results are shown in Table 3. All the gradient boosting trees models take the setup described in *Section Methods/Machine learning algorithms/Gradient boosting trees*.

In TopBP-ML(Complex), alpha complex is used to describe the arrangement of carbon and heavy atom networks, while Rips complex with different distance matrices is used to describe the protein-ligand interactions from the perspective of interaction distances and strength of electrostatics interactions. In detail, the features used are R-B0-I-C, R-B0-CI-S, A-B12-E-S as described in *Section Discussion/Complex based protein-ligand binding affinity prediction*, and those used in TopBP-ML(Ligand).

With the idea that a sequence of element combinations ordered by their importance in gradient boosting trees models can make an extra dimension of the description, we build a 2D convolutional neural network with one spatial dimension and one dimension of element combination. We combine this 2D CNN with a 1D CNN with the pairwise interaction inputs. For the construction of 2D input, the reader is referred to *Section Feature generation from topological invariants*. The 1D image-like inputs consist of two parts both generated by the counts in bins method described in *Section Feature generation from topological invariants*. For the 0th dimensional barcodes from interactive persistent homology of the 36 pairs of atom types ({C,N,O,S} from protein and {C,N,O,S,P,F,Cl,Br,I} from ligand), the interval [0, 50] Å is divided into equal length subintervals of length 0.25 Å. For the 0th dimensional barcodes from interactive persistent homology for electrostatics of the 50 pairs of atom types ({C,N,O,S,H} from protein and {C,N,O,S,P,F,Cl,Br,I,H} from ligand), the parameter interval of [0, 1] is divided into equal length subintervals of length 0.01. These two 1D image-like features have

Table 3. Pearson correlation coefficients (RMSE in kcal/mol) of different protein-ligand complex based approaches on PDBBind datasets.

Core set predictions					
Methods	v2007	v2013	v2015	v2016	Average
TopBP(Complex)	0.827 (1.93)	0.808 (1.95)	0.812 (1.92)	0.861 (1.65)	0.827 (1.86)
TopBP-ML(Complex)	0.818 (2.01)	0.804 (2.00)	0.797 (1.99)	0.848 (1.74)	0.817 (1.94)
TopBP-DL(Complex)	0.806 (1.95)	0.781 (1.98)	0.799 (1.91)	0.848 (1.64)	0.809 (1.87)
RF::VinaElem[a]	0.803 (1.94) [104]	0.752 (2.03) [105]	-	-	-
RI-Score [106] [b]	0.803 (1.99)[c]	-	0.762 (2.05)[c]	0.815 (1.85)	-
Refined set 5-fold cross validations					
Methods	v2007	v2013	v2015	v2016	Average
TopBP-ML(Complex)	0.752 (1.95)	0.768 (1.75)	0.781 (1.71)	0.785 (1.71)	0.771 (1.78)
RI-Score [106] [d]	-	-	-	0.747 (1.83)	-

Pearson correlation coefficients with RMSE (kcal/mol) in parentheses for predictions by different methods are listed. For the tests on core sets, the models are trained with the corresponding refined set minus the core set. Five-fold cross validation is done on refined sets. Results of TopBP-ML(Complex) are the medians of 50 repeated runs. For TopBP-DL(Complex), 100 independent models are generated at first. A consensus model is built by randomly choosing 50 models out of the 100, and this process is repeated 1000 times with the median reported. TopBP(Complex) is a consensus model combining TopBP-ML(Complex) and TopBP-DL(Complex). Each time, 50 single deep learning models are randomly selected to form TopBP-DL(Complex) and a TopBP-ML(Complex) model is randomly selected. The average of the two is taken as the output for TopBP(Complex). This process is repeated 1000 times with the median reported.

[a] The authors did not specify the number of repeated experiments and whether the reported performance is the best or the median of the experiments.

[b] The medians of Pearson correlation coefficient among the repeated experiments are listed.

[c] Only the best RMSEs among the repeated experiments are reported.

[d] The median results are reported.

https://doi.org/10.1371/journal.pcbi.1005929.t003

sizes 200 × 36 and 100 × 50. The network architecture is given in *Section Methods/Machine learning algorithms/Deep convolutional neural networks*.

The final model TopBP(Complex) takes the average of TopBP-ML(Complex) and TopBP-DL(Complex) with the assumption that the errors made by the two approaches are only partially correlated and thus averaging over them may cancel part of the errors. As a result, TopBP(Complex) delivers the best prediction performance on all four testing sets.

Structure-based virtual screening

In this section, we examine the performance of the proposed method for the main application in this paper, which is structure-based virtual screening which involves protein-compound complexes obtained by attempting to dock the candidates to the target proteins. The dataset is much larger than the two applications on protein-ligand binding affinity prediction which makes parameter tuning very time consuming. Therefore, the best performing procedures in ligand-based binding affinity prediction and protein-ligand-complex-based binding affinity prediction are applied in this virtual screening application.

DUD data set. The directory of useful decoys (DUD) [107, 108] is used to benchmark our topological approach for virtual screening. The DUD data set contains 40 protein targets from six classes, i.e., nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes. A total of 3,961 active ligand-target pairs were identified from literature. The number of ligands for each target ranges from tens to hundreds. At most 36 decoys were constructed for each ligand, from the ZINC database of commercially available compounds [109]. At the first step, the ZINC database of 3.5 million compounds was reduced to a database of 1.5 million compounds with similarity less than 0.9 to the ligands. The similarity was measured by Tanimoto coefficient on CACTVS type 2 fingerprints. The decoys were selected so that they possess similar physical properties to the ligands but have dissimilar molecular topology (topology in the sense of chemistry, not mathematical topology). A total of 32 physical properties were used including molecular weight, partition coefficient, and number of hydrogen bonding groups. This results in a total of 128,374 compound-target pairs. A discrepancy between calculated partial charges for the ligand and decoy sets was reported for the original release 2 of DUD datasets, which makes it trivial for virtual screening methods to distinguish between the two categories using those charges [110]. In this work, we use the data with recalculated Gasteiger charges for both ligand and decoy sets given by Armstrong *et al.* [110] in AutoDock Vina and our electrostatic persistence.

Data processing. In structure-based virtual screening, the possible complex structures of the target protein and the small molecule candidate are required. For the DUD dataset, the structures of the 40 protein targets, the ligands, and the decoys are given, and we generate the protein-compound complexes by using docking software. To this end, we first add missing atoms to the proteins by using the profix utility in Jackal software package [111]. The receptors and ligands or decoys are prepared using the scripts prepare_receptor4.py and prepare_ligand4.py provided by the AutoDockTools module in MGLTools package (version 1.5.6) [112]. The bounding box of the binding site is defined as a cube with edge size equal to 27 Å, centered at the geometric center of the crystal ligand. AutoDock Vina (version 1.1.2) [113] is used to dock the ligands or decoys to the receptors. The option exhaustiveness is set to 16 and all the other parameters are set to their default values. In each docking experiment, the pose having the lowest binding free energy reported by AutoDock Vina, is used by the reranking models.

Evaluation. Two measurements, the enrichment factor (EF) and the area under the receiver operating characteristic curve (AUC), are used to evaluate each method's ability of

discriminating actives from decoys. The AUC is defined as

$$\mathrm{AUC} = 1 - \frac{1}{N_a}\sum_{i=1}^{N_a}\frac{N_d^i}{N_d}, \tag{1}$$

where N_a is the number of active ligands, N_d is the total number of decoys, and N_d^i is the number of decoys that are higher ranked than the ith ligand [98]. An AUC value of 0.5 is the expected value of a random selection, whereas a perfect prediction results in an AUC of 1. The EF at x% denoted by $\mathrm{EF}_{x\%}$ evaluates the quality of the set of top x% ranked compounds, by comparing the percentage of actives in the top x% ranked compounds to the percentage of actives in the entire compound set. It is defined as

$$\mathrm{EF}_{x\%} = \frac{N_a^{x\%}}{N^{x\%}}\cdot\frac{N}{N_a}, \tag{2}$$

where $N_a^{x\%}$ is the number of active ligands in the top x% ranked compounds, $N^{x\%}$ is the number of top x% ranked compounds, N is the total number of compounds, and N_a is the total number of active ligands.

To evaluate the performance of various methods on the DUD data set, the entries associated with one protein target are used as the test set in the experiment on this protein target [98]. For the selection of the training set of a given protein target, we follow a procedure given in the literature [107], where the entries associated to the rest of the proteins, excluding those that are within the same class of the testing protein and those that have reported positive cross-enrichment with the testing protein, are taken as the training set. The 40 proteins are split into 6 classes [100]. A detailed list of proteins that are excluded from the training set of each protein is given in Table F in S1 Text.

Topology based machine learning models. Our topology based machine learning model, called *TopVS-ML*, relies on manually constructed features and utilizes ensemble of trees methods. For the complex with the small molecules (i.e., ligands and decoys) docked to the receptor, features R-B0-I-BP, R-B0-CI-S, and A-B12-E-S are used (see *Section Discussion/Complex based protein-ligand binding affinity prediction*), whereas features R-B012-M1-S and A-B012-E-S (see *Section Discussion/Ligand based protein-ligand binding affinity prediction*) are used for the small molecules. The gradient boosting trees method, random forest method, and extra trees method are employed as voters. The averaged probabilities output by the three methods are used for the classifier to decide the class of the testing samples. The modules *GradientBoostingClassifier*, *RandomForestClassifier*, and *ExtraTreesClassifier* in the scikit-learn package [114] (version 0.17.1) are used. The parameters for the three modules are listed in Table 4. TopVS-ML achieves a performance of AUC = 0.83, $\mathrm{EF}_{2\%}$ = 8.6, $\mathrm{EF}_{20\%}$ = 3.4. These values are the median values of 10 repeated experiments. Table G in S1 Text lists the result of each single experiment confirming that the performance is consistent across each repeated run.

Table 4. Parameters used in machine learning.

Method	Parameters
GBT	n = 2000, s = 0.5, cw = 100:1, lr = 0.01, mf = sqrt
RF	n = 2000, cw = balanced_subsample
ET	n = 2000, cw = balanced_subsample

The parameters used for the ensemble of trees methods while the other parameters are set to default. GBT: gradient boosting trees. RF: random forest. ET: extra trees. n: n_estimators. s: subsample. cw: class_weight. lr: learning_rate. mf: max_feature.

https://doi.org/10.1371/journal.pcbi.1005929.t004

Topology based deep learning model. Our topology based deep learning model, called *TopVS-DL*, relies on 1D image-like inputs for protein-compound complexes and manually constructed features for the compounds. The 2D representation used in binding affinity problem is not used here due to the intractable data size. The manually constructed features for the compounds are R-B012-M1-S and A-B012-E-S as described in *Section Discussion/Ligand based protein-ligand binding affinity prediction*. The 1D image-like inputs consisted of three parts are all generated by the counts in bins method described in *Section Feature generation from topological invariants*. (1) For the 0th dimensional barcodes from interactive persistent homology of the 36 pairs of atom types ({C, N, O, S} from protein and {C, N, O, S, P, F, Cl, Br, I} from ligand), the interval [0, 25] Å is divided into equal length subintervals of length 0.25 Å. The barcodes used here are identical to the barcodes in feature R-B0-I-BP. This results in a 1D image-like feature with size 100 × 36. (2) For the 0th dimensional barcodes from interactive persistent homology for electrostatics of the 50 pairs of atom types ({C, N, O, S, H} from protein and {C, N, O, S, P, F, Cl, Br, I, H} from ligand), the parameter interval of [0, 1] is divided into equal length subintervals of length 0.01. The barcodes used are identical to the barcodes in feature R-B0-CI-S. This results in a 1D image-like feature with size 100 × 50. (3) Alpha complex based persistent homology is applied to all carbon atoms and all heavy atoms. The computation is done on the complex as well as only the protein with a cutoff distance of 12 Å from the ligands. The interval [0, 12] Å is divided into equal length subintervals of length 0.125 Å. Counts in bins method is applied to the 0th, 1st, and 2nd dimensional barcodes. The features are generated for persistent homology computation of the complex and the protein. The features for the complex and the difference between the features for complex and protein are finally used. This results in a 1D image-like feature of size 96 × 32. The detailed network architecture is listed in *Section Methods/Machine learning algorithms/Deep convolutional neural networks*. A consensus model is constructed by taking the average over 25 single models trained independently. TopVS-DL achieves a performance of AUC = 0.81, $EF_{2\%}$ = 9.1, $EF_{20\%}$ = 3.2.

The final model. Same as the idea of taking the average output of different ensemble of trees models as the final output in TopVS-ML, we add TopVS-DL as another voter to TopVS-ML to construct a final model, called *TopVS*. Such consensus approach takes the average over different models with the hope that different models make partially uncorrelated errors which are possible to cancel out when averaged. The performance on each of 40 protein targets is reported in Table 5. We have also generated virtual screening results of AutoDock Vina (ADV) based on the computed binding free energy by ADV and compared them with those of the present TopVS in terms of enrichment factors and the areas under the receiver operating characteristic curve (AUC). A comparison of average AUC with those from a large number of methods is given in Table 6.

Discussion

Ligand based protein-ligand binding affinity prediction

We conduct several experiments on ligand based protein-ligand binding affinity prediction in this section which leads to the final models. To examine the strength and weakness of different sets of features and models, we first show a statistics fact of the S1322 data set of 7 protein clusters in Fig 2. The details of the S1322 data set is given in *Section Results/Ligand based protein-ligand binding affinity prediction*. All the gradient boosting trees models take the setup described in *Section Methods/Machine learning algorithms/Gradient boosting trees*.

Feature vectors for gradient boosting trees. In this test, Rips complex based and alpha complex based persistent homology computations up to 2nd dimension are performed for a variety of atom collections with different element types using the Euclidean metric and multi-

Table 5. Performance on each protein in DUD dataset.

Target	ADV			TopVS		
	$EF_{2\%}$	$EF_{20\%}$	AUC	$EF_{2\%}$	$EF_{20\%}$	AUC
ACE	4.1	1.4	0.42	5.1	3.1	**0.81**
AChE	4.7	2.8	**0.67**	1.4	1.9	0.65
ADA	0.0	0.4	0.49	7.8	4.5	**0.90**
ALR2	2.0	2.7	**0.74**	4.9	1.5	0.68
AmpC	2.4	0.2	0.34	0.0	1.0	**0.58**
AR	17.0	3.8	0.81	20.1	4.2	**0.90**
CDK2	9.0	2.4	0.64	7.6	4.1	**0.88**
COMT	13.1	1.4	0.56	17.4	2.9	**0.73**
COX1	9.9	2.8	0.76	11.8	3.6	**0.86**
COX2	20.7	3.9	0.86	23.3	4.9	**0.97**
DHFR	6.4	2.8	0.82	12.6	4.7	**0.96**
EGFr	3.4	1.6	0.63	16.4	4.8	**0.95**
$ER_{agonist}$	17.8	3.3	**0.84**	10.0	2.8	0.81
$ER_{antagonist}$	10.2	2.3	0.70	1.3	2.8	**0.83**
FGFr1	0.4	0.8	0.44	15.1	4.8	**0.95**
FXa	1.0	1.3	0.63	2.1	4.4	**0.89**
GART	0.0	1.9	**0.75**	2.6	0.7	0.48
GPB	0.0	0.9	0.48	1.4	1.5	**0.66**
GR	5.7	1.2	0.57	1.3	3.4	**0.84**
HIVPR	5.6	2.6	0.74	8.9	4.4	**0.91**
HIVRT	8.2	1.9	0.64	11.7	4.0	**0.88**
HMGR	0.0	0.9	0.53	14.4	5.0	**0.96**
HSP90	0.0	0.9	0.64	9.6	4.5	**0.93**
InhA	13.4	1.9	0.56	22.7	4.5	**0.95**
MR	16.7	4.0	0.82	0.0	4.3	**0.87**
NA	0.0	0.3	0.37	1.5	3.8	**0.87**
P38 MAP	1.4	1.7	0.59	18.4	4.5	**0.94**
PARP	4.2	2.7	0.71	0.0	1.7	0.71
PDE5	8.0	1.9	0.61	6.9	3.4	**0.86**
PDGFrb	3.5	0.5	0.32	26.5	4.9	**0.97**
PNP	0.0	0.7	0.59	7.9	4.3	**0.89**
PPARg	17.7	3.4	**0.82**	0.6	1.8	0.72
PR	1.9	1.1	0.52	9.4	4.1	**0.91**
RXRa	28.2	4.8	**0.95**	12.8	3.2	0.83
SAHH	10.4	3.0	0.80	4.5	3.9	**0.84**
SRC	5.6	2.3	0.71	24.6	4.9	**0.98**
thrombin	8.3	2.6	0.72	4.1	2.4	**0.79**
TK	0.0	0.9	0.56	6.9	2.5	**0.65**
trypsin	3.1	1.9	0.58	0.0	2.0	**0.78**
VEGFr2	10.2	2.2	0.63	24.9	4.7	**0.96**
Average	6.9	2.0	0.64	9.5	3.5	**0.84**

The median results of 10 repeated runs with different random seeds (for the TopVS-ML part) are reported. The best AUC in each row is marked in bold. The left block of AutoDock Vina (ADV) results are acquired from the ADV runs with the binding free energy reported by ADV.

https://doi.org/10.1371/journal.pcbi.1005929.t005

Table 6. AUC comparison of different methods on DUD dataset.

Method	AUC	Ref.
TopVS	0.84	
DeepVS-ADV	0.81	[98]
ICM[a]	0.79	[115]
NNScore1-ADV[b]	0.78	[97]
Glide SP[a]	0.77	[116]
DDFA-ALL	0.77	[100]
DDFA-RL	0.76	[100]
NNScore2-ADV[b]	0.76	[97]
DDFA-ADV	0.75	[100]
DeepVS-Dock	0.74	[98]
DDFA-AD4	0.74	[100]
Glide HTVS[b]	0.73	[97]
Surflex[a]	0.72	[116]
Glide HTVS	0.72	[116]
ICM	0.71	[115]
RAW-ALL	0.70	[100]
AutoDock Vina[b]	0.70	[97]
Surflex	0.66	[116]
Rosetta Ligand	0.65	[100]
AutoDock Vina	0.64	[100]
ICM	0.63	[116]
FlexX	0.61	[116]
Autodock4.2	0.60	[100]
PhDOCK	0.59	[116]
Dock4.0	0.55	[116]

[a]Tuned by expert knowledge.

[b]Determined using a different data set of decoys.

https://doi.org/10.1371/journal.pcbi.1005929.t006

level distance defined in Eq (3). Two types of features are generated and are denoted by F^C, which is a combination of F^C_b, F^C_d, and F^C_p, and F^S, which is a combination of F^S_b, F^S_d, and F^S_p. The construction of features F^C and F^S are described in *Section Feature generation from topological invariants*. For sets of the 0th dimensional bars, only F^C_d and F^S_d are computed. In each protein cluster, 10-fold or 5-fold cross validation is repeated 20 times for each subset of feature vectors depending on selected element type. The median Pearson correlation coefficients and the root-mean-square error (RMSE) in kcal/mol are reported. For Rips complex, both level 0 computation with distance matrix $\mathbf{M}$ and level 1 computation with distance matrix $\widetilde{\mathbf{M}}^1$ as defined in Eq (4) are performed. A comparison of these results is shown in S1 Text Table B. The results corresponding to alpha complex are shown in S1 Text Table A. The average performance for alpha complex and Rips complex has a Pearson correlation coefficient of 0.987.

Barcode space metrics for k-nearest neighbor regression. The barcodes generated using Rips complex with distance matrices $\mathbf{M}$ and $\widetilde{\mathbf{M}}^1$ are collected and the distance between each pair of barcodes are measured using the Wasserstein metric d^2. Leave-one-out prediction for every sample is performed with k-nearest neighbor regression with $k = 3$ within each protein cluster based on the Wasserstein metric. The results are shown in S1 Text Table C. The

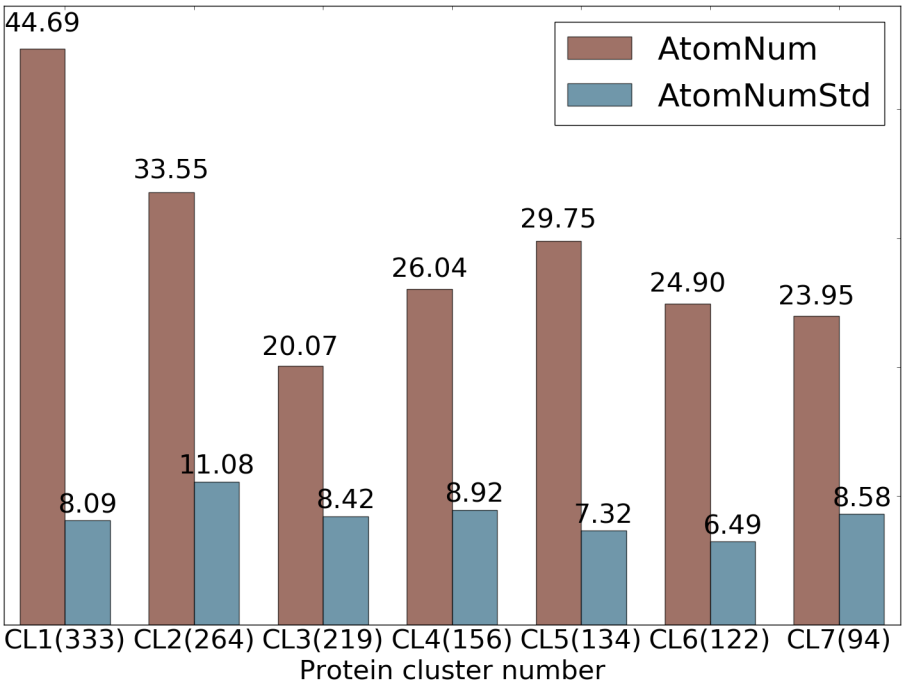

Fig 2. Statistics of ligands in 7 protein clusters in S1322 dataset. The average numbers of heavy atoms of a ligand in each protein cluster are shown in red and the standard deviations of number of heavy atoms across each protein cluster are shown in blue. The number of ligands in each cluster is given in parentheses.

https://doi.org/10.1371/journal.pcbi.1005929.g002

performance of the best performing and the worst performing protein clusters is shown in Fig 3. The better the performance, the closer the lines are to the semicircle.

The experiments done for this section are summarized in Table 7.

Performance of multi-component persistent homology. It can be noticed from Table 8 that topological features generated from barcode statistics typically outperform those created from counts in bins. R-B012-E-S-GBT and R-B012-M1-S-GBT perform similarly in the majority of the protein clusters whilst R-B012-M1-S-GBT which is based on $\widetilde{\mathbf{M}}^1$ significantly

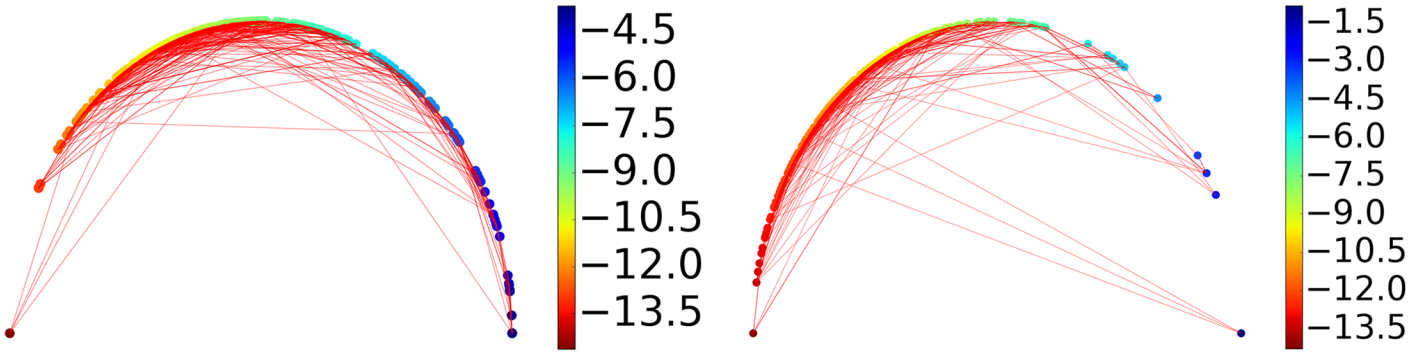

Fig 3. An illustration of similarities between ligands measured by their barcode space Wasserstein distances. Ligands are ordered according to their binding affinities and are represented as dots on the semicircle. Specifically, a sample of binding free energy x is plotted at the angle $\theta = \pi(E_{max} - x)/(E_{max} - E_{min})$ where E_{min} and E_{max} are the lowest and the highest energy in the dataset. Each dot is connected with two nearest neighbors based on their barcode space Wasserstein distances. An optimal prediction would be achieved if lines stay close to the semicircle. The majority of the connections stay near the boundary to the upper half sphere demonstrating that barcode space metric based Wasserstein distance measurement reflects the similarity in function, i.e., the binding affinity in this case. The protein clusters with the best and the worst performance are shown. Left: Protein cluster 2. Right: Protein cluster 3.

https://doi.org/10.1371/journal.pcbi.1005929.g003

Table 7. Experiments for ligand-based protein-ligand binding affinity prediction of 7 protein clusters and 1322 protein-ligand complexes.

Experiment	Description
A-B012-E-C-GBT	The barcodes are generated using alpha complex on different sets of atoms based on different element combinations. The features are constructed using the 0th, 1st, and 2nd dimensional barcodes following the *counts in bins* method with bins equally dividing the interval [0, 5]. Here 32 different element combinations are considered, including {C, N, O, S, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS, CNOSPFClBrI, H, CH, NH, OH, SH, CNH, COH, CSH, NOH, NSH, OSH, CNOH, CNSH, COSH, NOSH, CNOSH, CNOSPFClBrIH}. Gradient boosting trees (GBT) with the structured feature matrix are used for this computation.
A-B012-E-S-GBT	The barcodes same as those used in A-B012-E-C-GBT are used. Instead of *counts in bins*, the *Barcode statistics* method is used to generate features.
A-B012-E-SS-GBT	The barcodes same as those used in A-B012-E-C-GBT are used. The *persistence diagram slice and statistics* method is used to generate features. A uniform set of bins by dividing the interval [0, 5] into 10 equal length bins is used to slice birth, death, and persistence values.
R-B012-E-S-GBT	Barcodes are generated using Rips complex with Euclidean distances. The features are generated following the *barcode statistics* method. Here 36 element combinations are considered, i.e., {C, N, O, S, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS, CNOSPFClBrI, H, CH, NH, OH, SH, CNH, COH, CSH, NOH, NSH, OSH, CNOH, CNSH, COSH, NOSH, CNOSH, CNOSPFClBrIH, CCl, CClH, CBr, CBrH}.
R-B012-M1-S-GBT	The result is obtained with the same setup as R-B012-E-S-GBT except that the first level enrichment distance matrix $\widetilde{\mathbf{M}}^1$ is used instead of Euclidean distance.
R-B*n*-E-KNN	The *n*th dimensional barcodes from Rips complex computation with Euclidean distance are used. K-nearest neighbor (KNN) regression is performed with Wasserstein metric d^2. The leave-one-out validation is performed individually with each element combination and the average prediction of these element combinations is taken as the output result. The element combinations considered are {CNOS, CNOSPFClBrI, NOH, CNO, CNOSPFClBrIH}. These combinations are selected based on their performance in the gradient boosting trees experiments.
R-B*n*-M1-KNN	The result is obtained with the same setup as R-B*n*-E-KNN except that the distance matrix $\widetilde{\mathbf{M}}^n$ is used instead of Euclidean distance.

https://doi.org/10.1371/journal.pcbi.1005929.t007

outperforms R-B012-E-S-GBT which is based on Euclidean distance in protein cluster 3 and 6. To assess in what circumstances does the multi-level persistent homology improve the original persistent homology characterization of small molecules, we analyze the statistics of the size of ligands in Fig 2. It turns out that protein cluster 3 has the smallest average number of heavy atoms and protein cluster 6 has the smallest standard deviation of the number of heavy atoms. This observation partially answers the question that in the cases where the small molecules are relatively simple and are relatively of similar size, multi-level persistent homology is able to enrich the characterization of the small molecules which further improves the robustness of the model. Such enrichment or improvement over the original persistent homology approach is mainly realized in higher dimensional barcodes, i.e. the 1st and 2nd dimensions. In Table 8, the results with ID through 7 to 12 confirm that the 0th dimensional features from computation with $\widetilde{\mathbf{M}}^1$ are inferior to the results with Euclidean distance whilst the 1st and 2nd dimensional features based on $\widetilde{\mathbf{M}}^1$ outperforms the best result with Euclidean distance in most cases.

It is interesting to note that although Wasserstein metric based KNN methods are not as accurate as GBT approaches, the consensus result obtained by averaging over various predictions with Wasserstein metric on different sets of barcodes is quite accurate.

Robustness of topological learning models. Certain elements such as Br are very rare in the data sets studied in this work. Considering only the elements of high occurrence will not hurt the performance on the validations performed. However, omitting the low occurrence elements will sacrifice the capability of the model to handle new data in which such elements play an important role. Therefore, we decide to keep the rare elements that result in a large

Table 8. Performance of different approaches on the S1322 dataset.

ID	Experiments	CL 1 (333)	CL 2 (264)	CL 3 (219)	CL 4 (156)	CL 5 (134)	CL 6 (122)	CL 7 (94)	Average
1	A-B012-E-C-GBT	0.695(1.63)	0.836(1.18)	0.690(1.52)	0.642(1.38)	**0.840(1.30)**	0.647(1.65)	0.730(1.27)	0.726(1.42)
2	A-B012-E-S-GBT	0.695(1.63)	0.845(1.14)	0.678(1.54)	**0.692(1.31)**	0.828(1.35)	0.702(1.54)	0.739(1.25)	0.740(1.39)
3	A-B012-E-SS-GBT	0.704(1.62)	0.846(1.15)	0.681(1.53)	0.668(1.35)	0.834(1.34)	0.715(1.53)	0.741(1.25)	0.741(1.40)
4	R-B012-E-S-GBT	0.712(1.60)	0.837(1.17)	0.659(1.57)	0.683(1.32)	0.808(1.41)	0.635(1.67)	**0.757(1.22)**	0.727(1.42)
5	R-B012-M1-S-GBT	**0.716(1.59)**	0.836(1.17)	**0.706(1.48)**	0.672(1.34)	0.822(1.37)	**0.708(1.53)**	0.746(1.24)	0.744(1.39)
6	2+5	0.714(1.59)	**0.848(1.13)**	0.699(1.50)	**0.692(1.31)**	0.831(1.34)	**0.717(1.52)**	0.747(1.24)	**0.750(1.38)**
7	R-B0-E-KNN	0.648(1.73)	0.761(1.39)	0.544(1.76)	0.616(1.42)	0.700(1.70)	0.487(1.89)	0.641(1.43)	0.628(1.62)
8	R-B1-E-KNN	0.547(1.91)	0.684(1.55)	0.444(1.88)	0.536(1.52)	0.535(2.01)	0.634(1.67)	0.649(1.42)	0.576(1.71)
9	R-B2-E-KNN	0.474(2.01)	0.494(1.87)	0.202(2.14)	0.298(1.79)	0.126(2.49)	0.331(2.09)	0.609(1.47)	0.362(1.98)
10	R-B0-M1-KNN	0.581(1.85)	0.771(1.35)	0.516(1.80)	0.601(1.44)	0.672(1.76)	0.485(1.90)	0.644(1.43)	0.610(1.65)
11	R-B1-M1-KNN	0.663(1.70)	0.784(1.33)	0.652(1.59)	0.555(1.50)	0.786(1.49)	0.610(1.71)	0.731(1.30)	0.683(1.52)
12	R-B2-M1-KNN	0.675(1.67)	0.803(1.28)	0.577(1.72)	0.531(1.52)	0.655(1.81)	0.617(1.72)	0.648(1.42)	0.644(1.59)
13	Cons(7+8+9+10+11+12)	0.698(1.66)	0.817(1.28)	0.620(1.68)	0.645(1.41)	0.756(1.68)	0.658(1.68)	0.739(1.31)	0.705(1.49)
14	2+5 (5-fold)	0.713(1.60)	0.843(1.15)	0.693(1.51)	0.670(1.35)	0.831(1.34)	0.698(1.56)	0.737(1.26)	0.741(1.40)

Pearson correlation coefficients with RMSE (kcal/mol) in parentheses for binding affinity predictions on 7 protein clusters (CL) in S1322. On the title row, the numbers in parentheses denote the numbers of ligands in the cluster. The median results of 20 repeated runs are reported for the ensemble of trees based methods to account for randomness in the algorithm. For experimental labels, the first letter indicates the complex definition used, 'A' for alpha complex and 'R' for Rips complex. The second part starting with 'B' followed by the integers indicates the dimension of barcode used. The third part indicates the distance function used, 'E' for Euclidean and 'M1' for $\widetilde{\mathbf{M}}^1$. For row 1 through 5, the forth part shows the way of feature construction, 'C' for counts in bins and 'S' for barcode statistics. The last part indicates the regression technique used, 'GBT' for gradient boosting trees and 'KNN' for k-nearest neighbors. The detailed descriptions of the experiments are given in Table 7. Row 6 is the results using features of both row 2 and row 5. Row 13 is the consensus results by taking the average of the predictions by row 7 through row 12. Except for specified, all results are obtained from 10-fold cross validations.

https://doi.org/10.1371/journal.pcbi.1005929.t008

number of features and redundancy in features. For example, the element combinations CBrH and CH will probably deliver the same performance for most of the samples in the data sets studied in this work. To test whether this redundancy causes degenerated results of the model, the features of one element combination is added to the model at a step and the model is validated with an accumulation of the added features at each step. The performance of the model is measured with Pearson correlation coefficient and is plotted against number of element combinations involved in Fig 4. For most cases in Fig 4, the model is robust against the inclusion of more element combinations.

Complex based protein-ligand binding affinity prediction

Having demonstrated the representational power of the present topological learning method for characterizing small molecules, we further examine the method on the task of characterizing protein-ligand complex. Biologically, we consider the same task, i.e., the prediction of protein-ligand binding affinity, with a different approach that is based on the structural information of the protein-ligand complexes. Only gradient boosting trees and deep convolutional neural network algorithms are used in this section. All the gradient boosting trees models take the setup described in *Section Methods/Machine learning algorithms/Gradient boosting trees*.

In the present topological learning study, we use four versions of PDBBind core sets as our test sets. For each test set, the corresponding refined set, excluding the core set, is used as the training set.

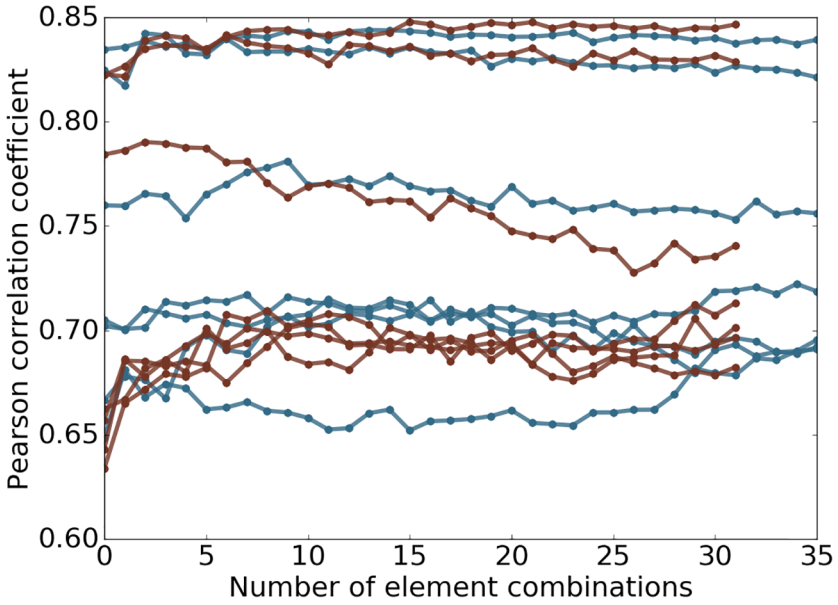

Fig 4. Plot of performance against number of element combinations used. The topological learning model performance against the number of element combinations involved in feature construction for 7 protein clusters in S1322. The horizontal axis corresponds to the number of element combinations used for the features. From left to right, one extra element combination is added at a step. The features are then used in gradient boosting trees method to test if the model is robust against redundant information. The results related to alpha complex are marked in red and Rips complex in blue. The median Pearson correlation coefficient between predicted and experimental results is reported of 10-fold cross-validation within each protein cluster repeated 20 times are reported.

https://doi.org/10.1371/journal.pcbi.1005929.g004

Groups of topological features and their performance in association with GBT. The experiments of protein-ligand-complex-based protein-ligand binding affinity prediction for the PDBBind datasets are summarized in Table 9.

Robustness of GBT algorithm against redundant element combination features and potential overfitting. It is intuitive that combinations of more than 2 element types are able to enrich the representation especially in the case of higher dimensional barcodes. However, the consideration of combination of more element types rapidly increases the dimension of feature space. In the high dimensional feature space, it is almost inevitable that there exists nonessential and redundant features. Additionally, the importance of a feature varies across different problems and data sets. Therefore, it is preferable to keep all the potentially important features in a general model which is expected to cover a wide range of situations. To test the robustness of the model against unimportant features, we select a total of 128 element combinations (i.e., all possible paired choices of one item from {C, N, O, CN, CO, NO, CNO, CNOS} in protein and another item from {C, N, O, S, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS, CNOSPFClBrI} in ligand). The 0th, 1st, and 2nd dimensional barcodes are computed for all combinations using alpha complex with Euclidean distance. Features are generated following the barcode statistics method.

A general model with all the features is generated in the first place. The element combinations are then sorted according to their importance scores in the general model. Starting from the most important element combination, one element combination is added to the feature vector each time and then the resulting feature vector is passed to the machine learning training and testing procedure. The order of adding element combinations is based on their importance scores and thus that a less important feature is added each step.

Table 9. Experiments for protein-ligand-complex-based protein-ligand binding affinity prediction for the PDBBind datasets.

Experiment	Description
R-B0-I-C	0th dimensional barcodes from Rips complex computation with interactive distance matrix based on Euclidean distance are used. Features are generated following *counts in bins* method with bins {[0, 2.5), [2.5, 3), [3, 3.5), [3.5, 4.5), [4.5, 6), [6, 12]}. Element combinations used are all possible paired choices of one item from {C, N, O, S, CN, CO, NO, CNO} in protein and another item from {C, N, O, S, P, F, Cl, Br, I, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS} in ligand, which result in a total of 160 combinations.
R-B0-I-BP	The persistent homology computation and feature generation is the same as R-B0-I-C. However, the element combinations used are all possible paired choices of one item from {C, N, O, S} in protein and another item from {C, N, O, S, P, F, Cl, Br, I} in ligand, which result in a total of 36 element combinations.
R-B0-CI-C	0th dimensional barcodes from Rips complex computation with interactive distance matrix based on the electrostatics correlation function defined in Eq (10) with the parameter $c = 100$. The features are generated following *counts in bins* method with bins {(0, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], (0.9, 1.0)}. The element combinations used are all possible paired choices of one item from {C, N, O, S, H} in protein and another item from {C, N, O, S, P, F, Cl, Br, I, H} in ligand, which result in a total of 50 element combinations.
R-B0-CI-B-S	The barcodes and element combinations are the same as those of R-B0-CI-B-C. The features are generated following the *barcode statistics* method.
A-B12-E-S	1st and 2nd dimensional barcodes from alpha complex computation with Euclidean distance are used. The element combinations considered are all heavy atoms and all carbon atoms. Features are generated following the *barcode statistics* method.

https://doi.org/10.1371/journal.pcbi.1005929.t009

Fig 5 depicts the changes of Pearson correlation coefficient and RMSE (kcal/mol) with respect to the increase of element combinations in predicting four PDBBind core sets. In all cases, the inclusion of top combinations can readily deliver very good models. The behavior of the present method in PDBBind v2007 is quite different from that in other data sets. The performance of the present method improves almost monotonically as the element combination increases. However, in other three cases, the improvement is unsteady. Nevertheless, the performance fluctuates within a small range, which indicates that the present method is reasonably stable against the increase in element combinations. From a different perspective, the increase in element combinations might lead to overfitting in machine learning. Since the model parameters are fixed before the experiments, it shows that GBT algorithms are not very sensitive to redundant features and are robust against overfitting.

Usefulness of more than 2 element types for interactive 0th dimensional barcodes. While using element combinations with more than 2 element types with higher dimensional barcodes enriches characterization of geometry, it remains to assess whether interactive 0th dimensional characterization will benefit from element combinations with more element types. As an example, we denote interactive 0th dimensional barcodes for carbon and nitrogen atoms from protein and oxygen atoms from ligand by $\mathbf{B}_{\mathrm{CN-O}}$, barcodes for carbon atoms from protein and oxygen atoms from ligand by $\mathbf{B}_{\mathrm{C-O}}$, and barcodes for nitrogen atoms from protein and oxygen atoms from ligand by $\mathbf{B}_{\mathrm{N-O}}$. In the case of persistent homology barcode representation, $\mathbf{B}_{\mathrm{CN-O}}$ is not strictly the union of $\mathbf{B}_{\mathrm{C-O}}$ and $\mathbf{B}_{\mathrm{N-O}}$. However $\mathbf{B}_{\mathrm{CN-O}}$ might be redundant to $\mathbf{B}_{\mathrm{C-O}}$ and $\mathbf{B}_{\mathrm{N-O}}$. To address this concern, we test features from interactive 0th dimensional barcodes with the 36 element combinations (i.e., {C, N, O, S} for protein and {C, N, O, S, P, F, Cl, Br, I} for ligand) and features for the 160 selected element combinations (i.e., {C, N, O, S, CN, CO, NO, CNO} for protein and {C, N, O, S, P, F, Cl, Br, I, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS} for ligand), which are listed as feature group 2 and feature group 1 in Table 10. In all the four cases, the features of the 36 combinations (feature group 2) slightly outperforms or performs as well as the features of the 160 combinations (feature group

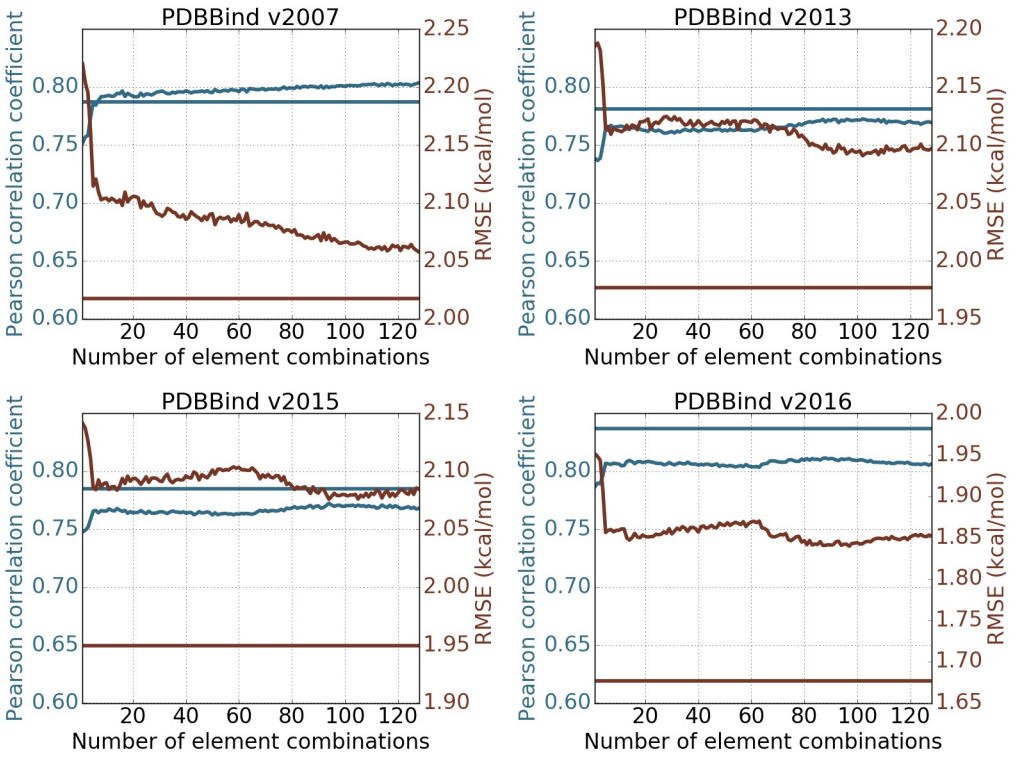

Fig 5. Feature robustness tests on PDBBind datasets. The performance of the topological learning model against the number of included element combinations for predicting on PDBBind core sets and training on PDBBind refined sets minus the core sets. The 1st and 2nd dimensional barcodes computed with alpha complex is used. Features are generated following *barcode statistics* method. Element combinations are all possible paired choices of one item from {C, N, O, CN, CO, NO, CNO, CNOS} in protein and another item from {C, N, O, S, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS, CNOSPFClBrI} in ligand, which result in 128 element combinations. The horizontal straight lines represents the performance of the 2D representation with deep convolutional neural network (row 10 in Table 10). The blue and red colors correspond to Pearson correlation coefficient and RMSE (kcal/mol) respectively. Each experiment is done by training on refined set minus the core set with the median result of 20 repeated runs reported.

https://doi.org/10.1371/journal.pcbi.1005929.g005

1) suggesting that element combinations with more than 2 element types are redundant to all the combinations with 2 element types in the case of interactive 0th dimensional characterization.

Importance of atomic charge in electrostatic persistence. In element specific persistent homology, atoms of different element types are characterized separately, which offers a rough and implicit description of the electrostatics of the system. However, such implicit treatment of electrostatics may lose important information because atoms behave differently at different oxidation states. Therefore, we explicitly embed atomic charges in interactive 0th dimensional barcodes as described in Eq (10). The resulting topological features are given in feature group 4 in Table 10. It can be seen from Table 10 that the combination of feature group 4 and the Euclidean distance based interactive 0th dimensional barcodes (listed as feature group 6 and 7) generally outperforms the results obtained with only Euclidean distance based features. This observation suggests that electrostatics play an important role and should be taken care of explicitly for the protein-ligand binding problem. Additionally, the inclusion of physical interactions in topological invariants opens a promising new direction in topological analysis.

Relevance of elements that are rare with respect to the data sets. Since the majority of the samples in both training and testing sets only contain atoms of element types, C, N, O, and H, the performance of the model on the samples with rare occurring elements with respect to

Table 10. Performance of different protein-ligand complex based approaches on the PDBBind datasets.

ID	Experiments	v2007	v2013	v2015	v2016	Average
1	R-B0-I-C	0.799 (2.01)	0.741 (2.14)	0.750 (2.11)	0.813 (1.82)	0.776 (2.02)
2	R-B0-I-BP	**0.816 (1.94)**	0.741 (2.13)	0.750 (2.10)	0.825 (1.78)	0.783 (1.99)
3	R-B0-CI-C	0.791 (2.05)	0.759 (2.10)	0.738 (2.13)	0.801 (1.87)	0.772 (2.04)
4	R-B0-CI-S	0.773 (2.10)	0.762 (2.12)	0.749 (2.13)	0.810 (1.86)	0.774 (2.05)
5	A-B12-E-S	0.736 (2.25)	0.709 (2.26)	0.695 (2.27)	0.752 (2.02)	0.723 (2.20)
6	1+4	0.815 (1.95)	0.780 (2.04)	0.774 (2.04)	0.833 (1.76)	0.801 (1.95)
7	2+4	0.806 (1.99)	0.787 (2.04)	0.770 (2.06)	0.834 (1.77)	0.799 (1.97)
8	1+4+5	0.810 (1.98)	0.792 (2.02)	0.786 (2.02)	0.831 (1.76)	0.805 (1.95)
9	2+4+5	0.802 (2.01)	**0.796** (2.02)	0.782 (2.04)	0.822 (1.79)	0.801 (1.97)
10	2D-CNN-Alpha	0.787 (2.02)	0.781 (**1.98**)	0.785 (1.95)	0.837 (1.68)	0.798 (1.91)
11	1D2D-CNN	0.806 (1.95)	0.781 (**1.98**)	**0.799 (1.91)**	**0.848 (1.64)**	**0.809 (1.87)**

Pearson correlation coefficients with RMSE (kcal/mol) in parentheses for predictions by various groups of features on the four PDBBind core sets. The training sets are the PDBBind refined sets minus the core sets of the same version year. Results of ensemble of trees based methods (rows 1 through 9) are the *median values* of 50 repeated runs to account for randomness in the algorithm. For the deep learning based methods (row 10 and 11), 100 independent models are generated in the first place. A consensus model is built by randomly choosing 50 models out of the 100, and the this process is repeated 1000 times with the median reported. The first letter indicates the definition of complex, 'A' for alpha complex and 'R' for Rips complex. The second part indicates the dimension of barcodes used. The third part indicates the distance function used, 'I' for $\widehat{\mathbf{M}}_{ij}$ defined in Eq (5), 'CI' for the one defined in Eq (10), and 'E' for Euclidean. The last part shows the way of feature construction, 'C' for counts in bins, 'S' for barcode statistics, and 'BP' for only pair of two single elements. The results reported in row 6 through 9 are obtained by combining the features of the rows with the corresponding numbers.

https://doi.org/10.1371/journal.pcbi.1005929.t010

data sets is hardly reflected by the overall performance statistics. For simplicity, we refer to such rarely occurring elements with respect to data sets simply by rarely occurring elements in the discussion follows. To assess the aspects of the model that potentially affect the performance on the samples containing rarely occurring elements, we picked the samples containing each rarely occurring element from the original testing set as a new testing set. Three experiments are carried out to address two questions: "Are the training samples containing the same rarely occurring element crucial?" and "Are features addressing the rarely occurring element important?". A short answer is yes to both according to the results shown in Fig 6. Specifically, for each rarely occurring element, the exclusion of samples containing this element in training set and the exclusion of features addressing this element will both cause degenerated results. It is also shown that the exclusion of samples of the rarely occurring element leads to much worse results. Since both modifications of the model deliver worse results, we conclude that including the samples in the training set with similar compositions to the test sample is crucial to the success of the model on this specific test sample. Even the inclusion of features of more element types or element combinations does not deliver better results in the general testing sets, such features should still be kept in the model in case that a sample with a similar element composition comes in as a test sample.

2D persistence for topological deep convolutional neural networks. Deep learning is potentially more powerful than many other machine learning algorithms when the data size is sufficiently large. In the present work, it is natural to construct a 2D topological representation by incorporating the element combination as an additional dimension, resulting in 16 channels as defined in *Section Feature generation from topological invariants*. Here 128 element combinations (i.e., all possible paired choices of one item from {C, N, O, CN, CO, NO, CNO, CNOS} in protein and another item from {C, N, O, S, CN, CO, CS, NO, NS, OS, CNO, CNS, COS, NOS, CNOS, CNOSPFClBrI} in ligand) are used for 2D analysis. The advantage of introducing this extra dimension with convolutional neural networks is to prevent unimportant

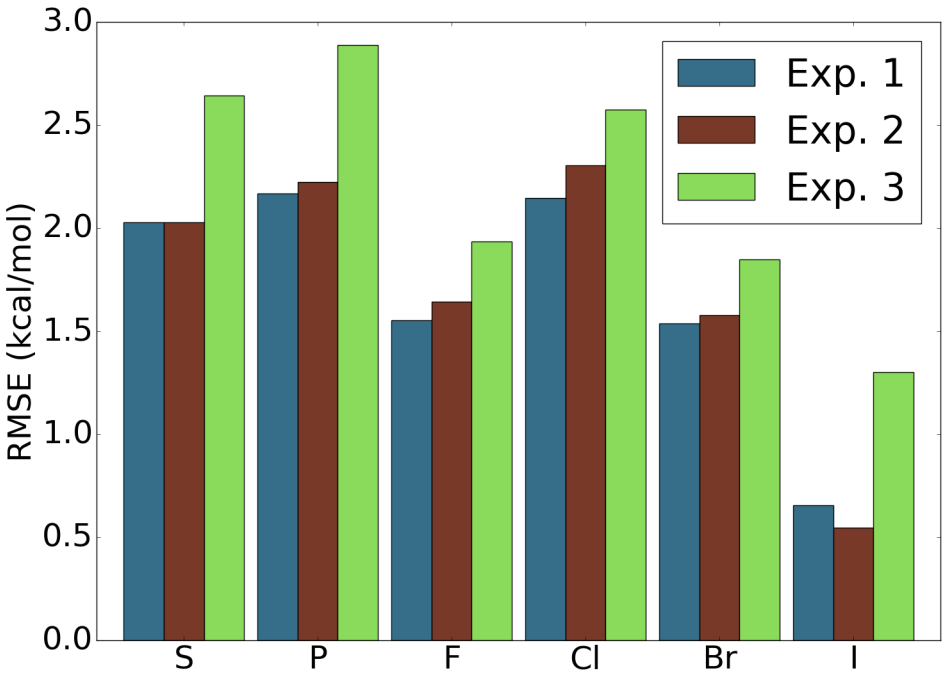

Fig 6. Assessment of performance of the model on samples with elements that are rare in the data sets. For the four data sets PDBBind v2007, v2013, v2015, and v2016 [99], and for each element, the testing set is the subset of the original core sets with only ligands that contain atoms of the particular element type. The features used are features with ID = 7 in Table 10. The reported RMSE is the average taken over the four data sets. Experiment 1: Training set is the original training set and all the features are used. Experiment 2: Training set is the original training set and only features that do not involve the particular element are used. Experiment 3: Training set is the original training set excluding the samples that contain atoms of the particular element type and all features are used. For most of the elements, experiment 1 achieves the best result and experiment 3 yields the worst performance.

https://doi.org/10.1371/journal.pcbi.1005929.g006

features from interacting with important ones at the lower levels of the model whilst generally unimportant features are still kept in the model in case that they are essential to specific problems or a certain portion of the data set. Fig 7 illustrates the mean value and the standard deviation of the PDBBind v2016 refined set. The existence of significant standard deviations for relatively unimportant element combinations indicates that these features might still contribute to the overall prediction.

As shown in Fig 5, for all the data sets except the PDBBind v2007 set, the 2D topological deep learning with convolutional neural networks performs significantly better. The inferior performance of convolutional neural networks in v2007 might be a result of the small data size. Note that v2007 training set has 1105 protein-ligand complexes, whereas other training sets have more than 2700 complexes. Consequently, topological deep convolutional neural networks are able to outperform the topological GBT algorithm in predicting v2013, v2015 and v2016 core sets. Indeed, topological deep convolutional neural networks have advantages in dealing with large data sets.

Structure-based virtual screening

In our final model TopVS reported in Table 6, we use topological descriptors of both protein-compound interactions and only the compounds (i.e., ligands and decoys) and take a consensus model on top of several ensemble of trees models and a deep learning model. We have also tested the behavior of our topological learning model TopVS-ML using either one of the

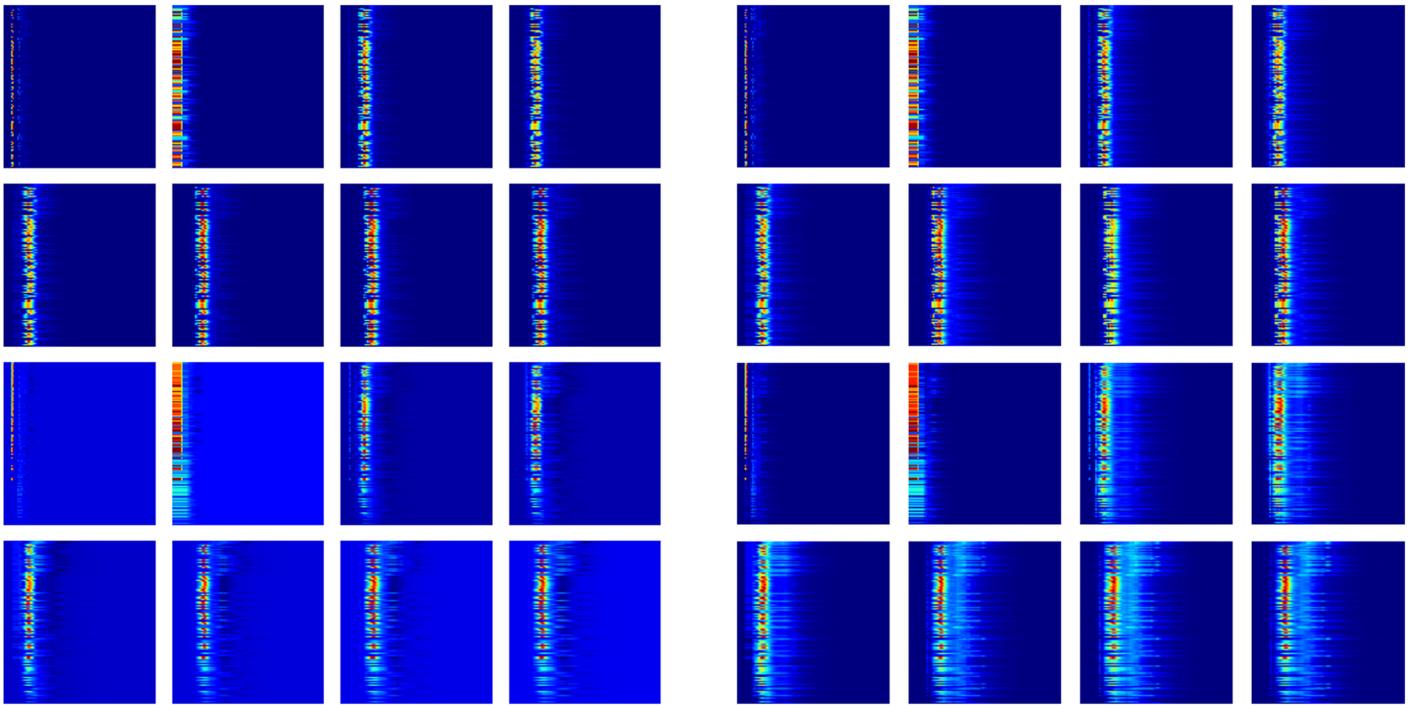

Fig 7. Heat map plot of the 16 channels. The mean value (left image) and the standard deviation (right image) of each digit over the PDBBind v2016 refined set are shown. The top 8 maps are for protein-ligand complex and the other 8 maps are for the difference between protein-ligand complex and protein only. For each map, the vertical axis is the element combinations ordered according to their importance and the horizontal axis is the dimension of spatial scales.

https://doi.org/10.1371/journal.pcbi.1005929.g007

aforementioned descriptions. The tests are done with TopVS-ML because that TopVS-DL is much more time consuming. When only topological descriptor of small molecules are used, which falls into the category of ligand-based virtual screening, an AUC of 0.81 is achieved. For the topological learning model using only the descriptions of protein-ligand interactions, an AUC of 0.77 is achieved. An AUC of 0.83 is obtained with a model combining both sets of descriptors which is better than each individual performance, suggesting that the two groups of descriptors are complementary to each other and are both important for achieving satisfactory results. The marginal improvement made by protein-compound complexes maybe due to the various docking quality. Similar situation was encountered by a deep learning method [98]. For the targets with high quality results by Autodock Vina (AUC of ADV > 0.8), the ligand-based features achieve an AUC of 0.81 and the complex-based features achieve an AUC of 0.86. On the other hand, for the targets with low quality results by Autodock Vina (AUC of ADV < 0.5), the ligand-based features achieve an AUC of 0.82 and the complex-based features achieve an AUC of 0.74. The results of these cases are listed in S1 Text, Tables H and I. This observation suggests that the performance of features describing the interactions and the geometry of protein-compounds complexes highly depends on the quality of docking results.

Our model with small molecular descriptors delivers an AUC of 0.81, which is comparably well to the other top performing methods. The performance of this model is also competitive in the regime of protein-ligand binding affinity prediction based on experimentally solved complex structures as is shown in *Section Discussion/Ligand based protein-ligand binding affinity prediction*. These results suggest that topology based small molecule characterization proposed in this work is potentially useful in other applications involving small molecules, such as predictions of toxicity, solubility and partition coefficient of small molecules.

Conclusion

Persistent homology is a relatively new branch of algebraic topology and is one of the main tools in topological data analysis. The topological simplification of biomolecular systems was a major motivation of the earlier persistent homology development [29, 36]. Persistent homology has been applied to computational biology [76, 77, 77–79], including our efforts [26, 87–91, 93]. However, the predictive power of primitive persistent homology was limited in early topological learning applications [92]. To address this challenge, we have recently introduced element specific persistent homology to retain chemical and biological information during the topological abstraction of biomolecules [14, 27, 94]. The resulting topological learning approach offers competitive predictions of protein-ligand binding affinity and mutation induced protein stability changes. However, persistent homology based approaches for small molecules have not been developed and its representability and predictive powers for the interaction of small molecules with macromolecules have not been extensively studied.

The present work further introduces multi-component persistent homology, multi-level persistent homology and electrostatic persistence for chemical and biological characterization, analysis and modeling. Multi-component persistent homology takes a combinatorial approach to create possible element specific topological representations. Multi-level persistent homology allows tailored topological descriptions of any desirable interaction in biomolecules which is especially useful for small molecules. Electrostatic persistence incorporates partial charges that are essential to biomolecules into topological invariants. These approaches are implemented via the appropriate construction of the distance matrix for filtration. The representation power and reduction power of multi-component persistent homology, multi-level persistent homology and electrostatic persistence are validated by two databases, namely PDBBind [99] and DUD [107, 108]. PDBBind involves more than 4,000 high quality protein-ligand complexes and DUD contains 128,374 compound-target pairs. Two classes of problems are used to test the proposed topological methods, including the prediction of protein-ligand binding affinities and the discrimination of active ligands from decoys (virtual screening). In both problems, we examine the representability of proposed topological learning methods on small molecules, which are somewhat more difficult to describe by persistent homology due to their chemical diversity, variability and sensitivity. Additionally, these methods are tested on their ability to handle the full protein-ligand complexes. Advanced machine learning methods, including Wasserstein metric based k-nearest neighbors (KNNs), gradient boosting trees (GBT), random forest (RF), extra trees (ET) and deep convolutional neural networks (CNN) are utilized in the present work to facilitate the proposed topological methods, rendering advanced topological learning algorithms for quantitative and qualitative biomolecular predictions. The thorough examination of the method on the prediction of binding affinity for experimentally solved protein-ligand complexes leads to a structure-based virtual screening method, TopVS, which outperforms other methods. The feature sets introduced in this work for small molecules and protein-ligand complexes can be extended to other applications such as 3D-structure based prediction of toxicity, solubility, and partition coefficient for small molecules and complex structure based prediction of protein-nucleic acid binding and protein-protein binding affinities.

Methods

Persistent homology

The concept of persistent homology is built on the mathematical concept of homology, which associates a sequence of algebraic objects, such as abelian groups, to topological spaces. For

discrete data such as atomic coordinates in biomolecules, algebraic groups can be defined via simplicial complexes, which are constructed from simplices, generalizations of the geometric notion of nodes, edges, triangles and tetrahedrons to arbitrarily high dimensions. Homology characterizes the topological connectivity of geometric objects in terms of topological invariants, i.e., Betti numbers, which are used to distinguish topological spaces by counting k-dimensional holes. Betti-0, Betti-1 and Betti-2, respectively, represent independent components, rings and cavities in a physical sense. In persistent homology, the generators in the homology groups are tracked along with a filtration parameter, such as the radius of a ball or the level set of a hypersurface function, that continuously varies over a range of values. Therefore, persistent homology is induced by the filtration. For a given biomolecule, the change and the persistence of topological invariants over the filtration offer a unique characterization. These concepts are very briefly discussed below. For more detailed theory and algorithms, the interested readers are referred to a book on computational topology [117].

Simplicial complex. A (geometric) *k-simplex* denoted σ^k is the convex hull of $k + 1$ affinely independent points in $\mathbb{R}^k$. The convex hull of each nonempty subset of the $k + 1$ points forms a subsimplex and is regarded as a *face* of σ^k. The points are also called *vertices* of σ^k.

A set of simplices K is a *simplicial complex* if all faces of any simplex in K are also in K and the intersection of any pair of simplices in K is either empty or a common face of the two simplices.

Homology. A *k-chain* of a simplicial complex K denoted by c_k is a formal sum of *k-simplices* in K. Here, we take the $\mathbb{Z}_2$ field for the coefficients of the formal sum. Under the addition operation of $\mathbb{Z}_2$, a group of *k-chains* is called a *chain group* and denoted $C_k(K)$ which has the basis as the set of *k-simplices* in K.

A *boundary operator* denoted by ∂_k: $C_k(K) \to C_{k-1}(K)$ maps a *k-chain* which is a linear combination of *k-simplices* to the same linear combination of the boundaries of the *k-simplices*. With a *k-simplex* $\sigma^k = [v_0, \ldots, v_k]$ where v_i are the vertices of σ^k, the *boundary operator* is defined as $\partial_k \sigma^k = \sum_{i=0}^{k} \sigma_i^{k-1}$, where σ_i^{k-1} is a *(k-1)-simplex* which is a face of σ^k with the ith vertex being absent. Since we are working with the $\mathbb{Z}_2$ coefficients, we omit the orientations of the simplices.

A *k-cycle* is a *k-chain* whose image under the *boundary operator* ∂_k is the empty set. The collection of all the *k-cycles* forms a group denoted $Z_k(K)$ which is the kernel of ∂_k: $C_k(K) \to C_{k-1}(K)$.

The image of ∂_{k+1}: $C_{k+1}(K) \to C_k(K)$ is called the boundary group and is denoted by $B_k(K)$. $B_k(K)$ is a subgroup of $Z_k(K)$ following the property of the *boundary operator* that $\partial_k \circ \partial_{k+1} = 0$.

The kth *homology group* is the quotient group defined as $H_k(K) = Z_k(K)/B_k(K)$. Its ranks are the Betti numbers of K and its generators (equivalence classes) are also of interest.

The kth *Betti number* β_k is defined and often computed as $\text{rank}H_k(K) = \text{rank}Z_k(K) - \text{rank}B_k(K)$. Intuitively, Betti numbers count the number of k-dimensional holes that can not be continuously deformed to each other. Analogous to the continuous case, in simplicial topology, two cycles (elements of $Z_k(K)$) that defer by the boundary of a chain (an element of $B_k(K)$) are considered to be able to deform continuously to each other and are thus representing the same element in $H_k(K)$.

Persistent homology. A *filtration* of a *simplicial complex* K is a nested sequence of subcomplexes of K such that $\emptyset = K^0 \subset K^1 \subset \ldots \subset K^m = K$. Each K^i is itself a *simplicial complex*. We are interested in tracking the birth and death of homology generators along filtration.

Given a *simplicial complex* K with its filtration, the *p-persistent kth homology group* of K^i is defined as $H_k^p(K^i) = Z_k(K^i)/(B_k(K^{i+p}) \cap Z_k(K^i))$. Intuitively, this records the homology classes of K^i that are persistent at least until K^{i+p}. Persistent homology allows us to not only compute

n-dimensional holes at a specific setup, but also compute the parameter values corresponding to the birth and death of the n-dimensional holes along the filtration.

A generator (equivalence class) in $H_k(K^i)$ which does not exist in $H_k(K^{i-1})$ or $H_k(K^j)$ and lasts until $H_k(K^{j-1})$, under the mappings along the sequence of homology groups induced by inclusion maps along the sequence of simplicial complexes, is associated with the interval $[x_i, x_j)$, where x_i and x_j are the filtration levels associated to K^i and K^j. A collection of these intervals tracks the appearing and disappearing of homology generators along the filtration process. Such collections of intervals can be visualized by stacking horizontal line segments (barcodes) or by plotting in a plane (persistence diagrams). The collection of intervals associated to the kth homology group is called the kth dimensional barcodes.

Simplicial complexes and filtration. Given a finite set of points X and a non-negative scale parameter r, the Vietoris-Rips complex and alpha complex are constructed as follows.

With a predefined distance function $d(\cdot, \cdot)$ in X, a subset X' of X forms a simplex if $d(x_i, x_j) \leq r$ for all $x_i, x_j \in X'$. The collection of all such simplices is the *Vietoris-Rips* complex of the finite metric space X with scale parameter r denoted by $Rips(X, r)$. It is obvious that $Rips(X, r) \subseteq Rips(X, r')$ for $r \leq r'$.

With $Alpha(X, r)$ being the *alpha complex* of X with the scale parameter r and given the Delaunay triangulation induced by the Voronoi diagram of X, a simplex in the Delaunay triangulation belongs to $Alpha(X, r)$ if all its *1-faces* (*1-simplex* as subset of the simplex) have length no greater than $2r$. Similar to Rips complex, alpha complex also has the property that $Alpha(X, r) \subseteq Alpha(X, r')$ for $r \leq r'$.

Biological considerations

The development of persistent homology was motivated by its potential in the dimensionality reduction, abstraction and simplification of biomolcular complexity [36]. In the early applications of persistent homology to biomolecules, emphasis was given on major or global features (long-persistent features) to derive descriptive tools. For example, persistent homology was used to identify the tunnel in a Gramicidin A channel [36] and to study membrane fusion [118]. For the predictive modeling of biomolecules, features of a wide range of scales might all be important to the target quantity [26]. At the global scale, the biomolecular conformation should be captured. At the intermediate scale, the smaller intra-domain cavities need to be identified. At the most local scale, the important substructures should be addressed, such as the pyrrolidine in the side chain of proline. These features of different scales can be reflected by barcodes with different centers and persistences. Therefore, applications in biomolecules can make a more exhaustive use of persistent homology [26, 87], compared to some other applications where only global features matter while most local features are mapped to noise. Earlier use of persistent homology was focused on qualitative analysis. Only recently had persistent homology been devised as a quantitative tool [26, 87]. While the aforementioned applications are descriptive and regression based analysis, we have also applied persistent homology to predictive modeling of biomolecules [92]. However, biomolecules are both structurally and biologically complex. Their geometric and biological complexities include covalent bonds, non-covalent interactions, effects of chirality, cis and trans distinctions, multi-leveled protein structures, and protein-ligand and protein-nucleic acid complexes. Covering a large range of spatial scales is not enough for a powerful model. The biological details should also be explored. We address the underlying biology and physics by modifying the distance function and selecting various sets of atoms according to element types, to describe different interactions. Some biological considerations are discussed in this section.

Covalent bonds. Covalent bonds are formed via shared electron pairs or bonding pairs. The lengths and the number of covalent bonds can be easily detected from 0th dimensional barcodes. For macromolecules, the same type of covalent bonds have very similar bond lengths and thus 0th dimensional barcode patterns.

Non-covalent interactions. Non-covalent interactions play a critical role in maintaining the 3D structure of biomolecules and mediating chemical and biological processes, such as solvation, binding, protein-DNA specification, molecular self-assembly, etc. Physically, non-covalent interactions are due to electrostatic, van der Waals forces, hydrogen bonds, π-effects, hydrophobic effects, etc. The ability to characterize non-covalent interactions is an essential task in any methodological development. The 1st and 2nd dimensional barcodes are suitable for the characterization of the arrangement of such interactions in a larger scale. Additionally, we propose multi-level persistence and electrostatic persistence to reveal local and pairwise non-covalent interactions via 0th dimensional barcodes as well.

Chirality, cis effect and trans effect. Chirality, cis and trans effects are geometric properties of many molecules. Among them, chirality is a symmetry property such that a chiral molecule cannot be superposed on its mirror image. Cis and trans effects are due to molecular steric and electronic effects. Chirality, cis and trans effects often play a role in molecular kinetics, activity and catalysis, and thus their characterization is an important issue in developing topological methods. These effects should be reflected from barcodes of various dimensions.

Multi-leveled protein structures. Protein structures are typically described in terms of primary, secondary, tertiary and quaternary levels. The protein primary structure is the linear sequence of amino acids in the polypeptide chain. Protein secondary structure refers to the local 3D structure of protein segments containing mainly α-helix and β-sheets, which are highly regular and can be easily detected by distinct Frenet-Serret frames. A tertiary structure refers to the 3D structure of a single polypeptide chain. Its formation involves various non-covalent and covalent interactions including salt bridges, hydrophobic effects, and often disulfide bonds. A quaternary structure refers to the aggregation of two or more individual folded protein subunits into a 3D multi-subunit complex. Protein structures are further complicated by its functional domains, motifs, and particular folds. The protein structural diversity and complexity result in the challenge and opportunity for methodological developments. Various persistent homology techniques, including multi-component, multi-level, multi-dimensional [119], multi-resolution [90], electrostatic, and interactive [27] persistent homologies have been designed either in our earlier work or in this paper for protein structural diversity and complexity.

Protein-ligand, protein-protein, and protein-nucleic acid complexes. Topological characterization of proteins is further complicated by protein interactions or binding with ligands (drugs), other proteins, DNA and/or RNA molecules. Although a normal protein involves only carbon (C), hydrogen (H), nitrogen (N), oxygen (O) and sulfur (S) atoms, its protein-ligand complexes bring a variety of other elements into the play, including, phosphorus (P), fluorine (F), chlorine (Cl), Bromine (Br), iodine (I), and many important biometals, such as calcium (Ca), potassium (K) sodium (Na), iron (Fe), copper (Cu), cobalt (Co), zinc (Zn), manganese (Mn), chromium (Cr), vanadium (V), tin (Sn), and molybdenum (Mo). Each biological element has important biological functions and its presence in biomolecules should be treated uniformly as a set of points in the point cloud data. The interaction of protein and nucleic acids can be very intricate. Qualitatively, multiscale and multi-resolution persistent homology demonstrates interesting features in 3D DNA structures [89]. Typically, 3D RNA structures are more flexible and difficult to extract topological patterns. Interactive persistent homology, element specific persistent homology and binned representation for persistent homology

outputs were designed to deal with interactions between protein-ligand, protein-protein, and protein-nucleic acid complexes [14, 27, 94]. These approaches worked well in protein-mutation site interactions [14]. Additionally, multi-level persistent homology and electrostatic persistence proposed in this work are useful tools to describe some other specific interactions.

Element specific persistent homology

One important issue is how to protect chemical and biological information during the topological simplification. As mentioned earlier, one should not treat different types of atoms as homogeneous points in a point cloud data. To this end, element specific persistent homology or multi-component persistent homology has been proposed to retain biological information in topological analysis [14, 27, 94]. The element selection is similar to a predefined vertex color configuration for graphs.

When all atoms are passed to persistent homology algorithms, the information extracted mainly reflects the overall geometric arrangement of a biomoelcule at different spatial scales. By passing only atoms of certain element types or of certain roles to the persistent homology analysis, different types of interactions or geometric arrangements can be revealed. In protein-ligand binding modeling, the selection of all carbon atoms characterizes the hydrophobic interaction network whilst the selection of all nitrogen and/or oxygen atoms characterizes hydrophilic network and the network of potential hydrogen bonds. In the protein structural analysis, computation on all atoms can identify geometric voids inside the protein which may suggest structural instability and computation on only C_α atoms reveals the overall structure of amino acid backbones. In addition, combination of various selections of atoms based on element types provides very detailed description of the biomolecular system and the hidden relationships from the structure to function can then be learned by machine learning algorithms. This may lead to the discovery of important interactions not realized as *a prior*. This can be realized by passing the set of atoms of the selected element types to the persistent homology computation. This concept is used with the various definitions of distance matrix discussed as follows.

Distance matrix induced persistent homology

Biomolecular systems are not only complex in geometry, but also in chemistry and biology. To effectively describe complex biomolecular systems, it is necessary to modify the filtration process. There are three commonly used filtrations, namely, radius filtration, distance matrix filtration, and density filtration, for biomolecules [26, 90]. A distance matrix defined with smoothed cutoff functions was proposed in our earlier work to deal with interactions within a spatial scale of interest in biomolecules [26]. In the present work, we introduce more distance matrices to enhance the representational power of persistent homology and to cover some important interactions that were not covered in our earlier works. The distance matrices can be used with a more abstract construction of simplicial complexes, such as Vietoris-Rips complex.

Multi-level persistent homology. Small molecules such as ligands in protein-ligand complexes usually contain fewer atoms than large biomolecules such as proteins. Bonded atoms stay closer than non-bonded ones in most cases. As a result, the collection of 0th dimensional bars will mostly provide the information about the length of covalent bonds and the higher dimension barcodes will most likely be very sparse. It is difficult to capture non covalent bond interactions among atoms especially hydrogen bonds and van der Waals pairwise interactions in 0th dimensional barcodes. In order to describe non covalent interactions, we propose multi-level persistent homology, by simply modifying the distance matrix, similar to the idea

of modifying distance matrix to emphasize on the interactions between protein and ligand [27]. Given the original distance matrix $\mathbf{M} = (d_{ij})$ with $1 \leq i, j \leq N$, the modified distance matrix is defined as

$$\widetilde{\mathbf{M}}_{ij} = \begin{cases} d_{\infty}, & \text{if atoms i and j are bonded,} \\ \mathbf{M}_{ij}, & \text{otherwise,} \end{cases} \tag{3}$$

where d_{∞} is a large number which is set to be greater than the upper limit of the filtration value chosen by a persistent homology algorithm. Note that this matrix may fail to satisfy triangle inequality whilst still satisfies the construction principle of Rips complex.

The present multi-level persistent homology is able to describe any selected interactions of interest and delivers two benefits in characterizing biomolecules. Firstly, the pairwise non-covalent interactions can be reflected by the 0th dimensional barcodes. Secondly, such treatment generates more higher dimensional barcodes and the small structural fluctuation among different conformations of the same molecule can be captured. The persistent barcode representation of the molecule can be significantly enriched to better distinguish between different molecular structures and isomers. As an illustration, we take the ligand from the protein-ligand complex with PDB code "1BCD" which only has 10 atoms. A different conformation of the ligand is generated by using the Frog2 web server [120]. The persistent barcodes generated using Rips complex with the distance matrices $\mathbf{M}$ are identical and only have 0th dimensional bars due to the simple structure. In this case, the 0th dimensional bars only reflect the length of each bond and therefore fail to distinguish the two slightly different conformations of the same molecule. However, when the modified distance matrices $\widetilde{\mathbf{M}}$ are employed, the barcode representation is significantly enriched and is able to capture the tiny structural perturbation between the conformations. An illustration of the outcome from the modified distance matrix $\widetilde{\mathbf{M}}$ is shown in Fig 8. A general nth level persistence characterization of molecules can be

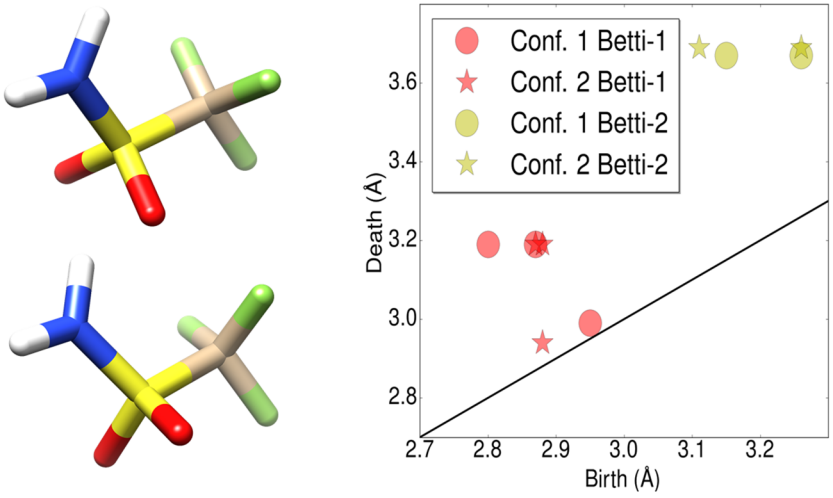

Fig 8. Multi-level persistent homology on simple small molecules. Illustration of representation ability of $\widetilde{\mathbf{M}}$ in reflecting structural perturbations among conformations of the same molecule. Left: The structural alignment of two conformations of the ligand in protein-ligand complex (PDB:1BCD). Right: The persistence diagram showing the 1st and 2nd dimensional results generated using Rips complex with $\widetilde{\mathbf{M}}$ for two conformations. It is worth noticing that the barcodes generated using Rips complex with $\mathbf{M}$ are identical for the two conformations.

https://doi.org/10.1371/journal.pcbi.1005929.g008

obtained with the distance matrix $\widetilde{\mathbf{M}}^n$ as,

$$\widetilde{\mathbf{M}}^n_{ij} = \begin{cases} d_\infty, D(i,j) \leq n \\ \mathbf{M}_{ij}, \text{ otherwise,} \end{cases} \tag{4}$$

where $D(i, j)$ is the smallest number of bonds to travel from atom i to atom j and d_∞ is some number greater than the upper limit of the filtration value.

Interactive persistent homology. In protein-ligand binding analysis and analysis involving interactions, we are interested in the change of topological invariants induced by interactions that are caused by binding or other processes. Similar to the idea of multi-level persistent homology, we can design a distance matrix to focus on the interactions of interest. For a set of atoms, $A = A_1 \cup A_2$ with $A_1 \cap A_2 = \emptyset$ where only interactions between atoms from A_1 and atoms from A_2 are of interest [27]. The interactive distance matrix $\widehat{\mathbf{M}}$ is defined as

$$\widehat{\mathbf{M}}_{ij} = \begin{cases} \mathbf{M}_{ij}, \text{ if } a_i \in A_1, \ a_j \in A_2 \text{ or } a_i \in A_2, \ a_j \in A_1, \\ d_\infty, \text{ otherwise,} \end{cases} \tag{5}$$

where $\mathbf{M}$ is the original distance matrix induced from Euclidean metrics or other correlation function based distances, a_i and a_j are atoms i and j, and d_∞ is a number greater than the upper limit of the filtration value. In applications, A_1 and A_2 can be respectively a set of atoms of the protein and a set of atoms of the ligand in a protein-ligand complex. In this case, the characterization of interactions between ligand and protein is an important task. In the modeling of point mutation induced protein stability changes, A_1 could be the set of atoms at the mutation site and A_2 could be the set of atoms of surrounding residues close to the mutation site. Similar treatment can be used for protein-protein and protein-nucleic acid interactions.

Correlation function based persistent homology. For biomolecules, the interaction strength between pair of atoms usually does not align linearly to their Euclidean distances. For example, van der Waals interaction is often described by the Lennard-Jones potential. Therefore, kernel function filtration can be used to emphasize certain geometric scales. Correlation function based filtration matrix was introduced in our earlier work [26]:

$$\bar{\mathbf{M}}_{ij} = 1 - \Phi(d_{ij}, \eta_{ij}), \tag{6}$$

where $\Phi(d_{ij}, \eta_{ij})$ is a radial basis function and η_{ij} is a scale parameter. This filtration can be incorporated in the element specific persistent homology

$$\grave{\mathbf{M}}_{ij} = \begin{cases} d_\infty, \text{ if atom i or atom j } \in \mathcal{U}, \\ 1 - \Phi(d_{ij}, \eta_{ij}), \text{ otherwise.} \end{cases} \tag{7}$$

Additionally, one can simultaneously use two or more correlation functions characterized by different scales to generate a multiscale representation of biomolecules [106].

Flexibility and rigidity index based filtration matrix. One form of the correlation function based filtration matrix is constructed by flexibility and rigidity index. In this case, the Lorentz function is used in Eq (7)

$$\Phi(d_{ij}; \eta_{ij}, \nu) = \frac{1}{1 + \left(\frac{d_{ij}}{\eta_{ij}}\right)^\nu}, \tag{8}$$

where d_{ij} is the Euclidean distance between point i and point j and η_{ij} is a parameter

controlling the scale and is related to radius of two atoms. When distance matrices based on such correlation functions are used, patterns at different spatial scales can be addressed separately by altering the scale parameter η_{ij}. Note that the rigidity index is given by [121]

$$\mu_i = \sum_j \Phi(d_{ij}; \eta_{ij}, \nu). \tag{9}$$

This expression is closely related to the rigidity density based volumetric filtration [90].

Electrostatic persistence. Electrostatic effects are some of the most important effects in biomolecular structure, function, and dynamics. The embedding of electrostatics in topological invariants is of particular interest and can be very useful in describing highly charged biomolecules such as nucleic acids and their complexes. We introduce electrostatics interaction induced distance functions in Eq (10) to address the electrostatic interactions among charged atoms. The abstract distance between two charged particles are rescaled according to their charges and their geometric distance, and is modeled as

$$\Phi(d_{ij}, q_i, q_j; c) = \frac{1}{1 + \exp(-cq_iq_j/d_{ij})}, \tag{10}$$

where d_{ij} is the distance between the two atoms, q_i and q_j are the partial charges of the two atoms, and c is a nonzero tunable parameter. c is set to a positive number if opposite-sign charge interactions are to be addressed and is set to a negative number if same-sign charge interactions are of interest. The form of the function is adopted from sigmoid function which is widely used as an activation function in artificial neural networks. Such function regularizes the input signal to the [0, 1] interval. Other functions can be similarly used. This formulation can be extended to systems with dipole or higher order multipole approximations to electron density. The weak interactions due to long distances or neutral charges result in correlation values close to 0.5. When $c > 0$, the repulsive interaction and attractive interaction deliver the correlation values in (0.5, 1) and (0, 0.5) respectively. The distances induced by $\Phi(d_{ij}, q_i, q_j; c)$ are used to characterize electrostatic effects. The parameter c is rather physical but chosen to effectively spread the computed values over the (0, 1) interval so that the results can be used by machine learning methods. Another simple choice of charge correlation functions is

$$\Phi(d_{ij}, \eta_{ij}, q_i, q_j) = q_iq_j \exp(-d_{ij}/\eta_{ij}).$$

However, this choice will lead to a different filtration domain. Additionally, a charge density can be constructed

$$\mu^c(\mathbf{r}) = \sum_j q_j \exp(-\parallel \mathbf{r} - \mathbf{r}_j \parallel /\eta_j), \tag{11}$$

where $\mathbf{r}$ is a position vector, $\|\mathbf{r} - \mathbf{r}_j\|$ is the Euclidean distance between $\mathbf{r}$ and jth atom position $\mathbf{r}_j$ and η_j is a scale parameter. Eq (11) can be used for electrostatic filtration as well. In this case, the filtration parameter can be the charge density value and cubical complex based filtration can be used.

Multi-component persistent homology. Multicomponent persistent homology refers to the construction of multiple persistent homology components from a given object to describe its properties. Obviously, element specific persistent homology leads to multi-component persistent homology. Nevertheless, in element specific persistent homology, the emphasis is given to the appropriate selection of important elements for describing certain biological properties or functions. For example, in biological context, electronegative atoms are selected for describing hydrogen bond interactions, polar atoms are selected for describing hydrophilic

interactions, and carbon atoms are selected for describing hydrophobic interactions. Note that in chemical context, an atom may have many sharply different chemical and physical properties, depending on its oxidation states. Whereas, in multicomponent persistent homology, the emphasis is placed on the systematic generation of topological invariants from different combinatorial possibilities and the construction of 2D or high-dimensional persistent maps for deep convolutional neural networks.

Feature generation from topological invariants

Barcode representation of topological invariants offers a visualization of persistent homology analysis. In machine learning analysis, we convert the barcode representation of topological invariants into structured feature arrays for machine learning. To this end, we introduce two methods, i.e., counts in bins, barcode statistics, and persistence diagram slice and statistics, to generate feature vectors from sets of barcodes. These methods are discussed below. Python code is given in S1 Code for the generation of features used in the final models in the Results section.

Counts in bins. For a given set of atoms A, we denote its barcodes as $\mathbf{B} = \{I_\alpha\}_{\alpha \in A}$ and represent each bar by an interval $I_\alpha = [b_\alpha, d_\alpha]$, where b_α and d_α are respectively the birth and death positions on the filtration axis. The length of each bar, or the persistence of topological invariant is given by $p_\alpha = d_\alpha - b_\alpha$. To locate the position of all bars and persistences, we further split the set of barcodes on the filtration axis into a predefined collection of N bins $\mathbf{Bin} = \{\mathrm{Bin}_i\}_{i=1}^N$ with $\mathrm{Bin}_i = [l_i, r_i]$, where l_i and r_i are the left and the right positions of the ith bin. We generate features by counting the numbers of births, deaths, and persistences in each bin, which leads to three counting feature vectors, namely, counts of birth F_b^C, death F_d^C, and persistence F_p^C,

$$\begin{aligned} F_{b,i}^C(\mathbf{B}) &= \| \{[b_\alpha, d_\alpha] \in \mathbf{B} | l_i \le b_\alpha \le r_i\} \|, 1 \le i \le N, \\ F_{d,i}^C(\mathbf{B}) &= \| \{[b_\alpha, d_\alpha] \in \mathbf{B} | l_i \le d_\alpha \le r_i\} \|, 1 \le i \le N, \\ F_{p,i}^C(\mathbf{B}) &= \| \{[b_\alpha, d_\alpha] \in \mathbf{B} | b_\alpha \le r_i \text{ or } l_i \le d_\alpha\} \|, 1 \le i \le N, \end{aligned} \tag{12}$$

where $\| \cdot \|$ is the cardinality of a set. Note that the above discussion should be applied to three topological dimensions, i.e., barcodes of the 0th dimension ($\mathbf{B^0}$), 1st dimension ($\mathbf{B^1}$) and 2nd dimension ($\mathbf{B^2}$). In general, this approach enables the description of bond lengths, including the length of non-covalent interactions, in biomolecules and was referred to as binned persistent homology in our earlier work [14, 27, 94].

Barcode statistics. Another method of feature vector generation from a set of barcodes is to extract important statistics of barcode collections such as maximum values and standard deviations. Given a set of bars $\mathbf{B} = \{[b_\alpha, d_\alpha]\}_{\alpha \in A}$, we define sets of **Birth** $= \{b_\alpha\}_{\alpha \in A}$, **Death** $= \{d_\alpha\}_{\alpha \in A}$, and **Persistence** $= \{d_\alpha - b_\alpha\}_{\alpha \in A}$. Three statistic feature vectors F_b^S, F_d^S, and F_p^S can then be generated in the sense of the statistics of the collection of barcodes. For example, F_b^S consists of avg(**Birth**), std(**Birth**), max(**Birth**), min(**Birth**), sum(**Birth**), and cnt(**Birth**), where avg(·) is the average value of a set of numbers, std(·) is the standard deviation of a set of numbers, max(·) and min(·) are maximum and minimum values in a set of numbers, sum(·) is the summation of elements in a set of numbers, and cnt(·) is the count of elements in a set. The generation of F_d^S is the same by examining the set **Death**. F_p^S contains the same information with two extra terms, the birth and death values of the longest bar. Statistics feature vectors are collected from barcodes of three topological dimensions, i.e., the 0th, 1st, and 2nd dimensions.

Persistence diagram slice and statistics. A more thorough description of sets of barcodes is to first divide the sets into subsets and extract features analogously to the *barcode statistics*

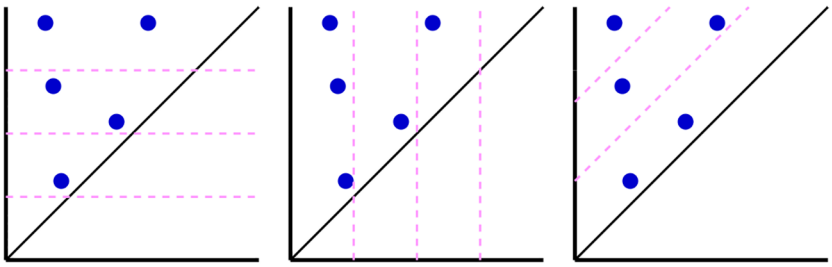

Fig 9. Illustration of dividing set of barcodes into subsets. The barcodes are plotted as persistence diagrams with the horizontal axis being birth and the vertical axis being death. From left to right, the subsets are generated according to the slicing of death, birth, and persistence values.

https://doi.org/10.1371/journal.pcbi.1005929.g009

method. As shown in Fig 9, a persistence diagram can be divided into slices in different directions. The barcodes that fall in each slice form a subset. Each subset is described in terms of feature vector by using the *barcode statistics* method. When the persistence diagram is sliced horizontally, members in each subset have similar death values and the barcode statistics feature vector is generated for the set of birth values. Similarly, members in each subset have similar birth values if the persistence diagram is sliced vertically, and the barcode statistics feature vector is generated for the set of death values. The barcode statistics feature vectors are generate for both set of birth values and set of death values if the persistence diagram is sliced diagonally, where members in each subset have similar persistence. This type of feature vector generation describes the set of barcodes in more detail but will produce longer feature vectors.

2D representation. The construction of multi-dimensional persistence is an interesting topic in persistent homology. In general, it is believed that multi-dimensional persistence has better representational power for complex systems described by multiple parameters [43]. Although multidimensional persistence is hard to compute, one can compute persistence for one parameter while fixing the rest of the parameters to a sequence of fixed values. In the case where there are two parameters, a bifiltration can be done by taking turns to fix one parameter to a sequence of fixed values while computing persistence for the other parameter. For example, one can take a sequence of resolutions and compute persistence for distance with each fixed resolution. The sequence of outputs can be stacked to form a multidimensional representation [119].

Computing persistence multiple times and stacking the results is especially useful when the parameters that are not chosen to be the filtration parameter are naturally discrete with underlying orders. For example, the multi-component or element specific persistent homology will result in many persistent homology computations over different selections of atoms. These results can be ordered by the percentage of atoms used of the whole molecule or by their importance scores in classical machine learning methods. Also, multiple underlying dimensions exist in the element specific persistent homology characterization of molecules. This property enables 2D or 3D topological representation of molecules. Based on the observation that the performance of the predictor degenerates when too many element combinations are used, we order the element combinations according to their individual performance on the task using methods of ensemble of trees. Combining the dimension of spatial scale and dimension of element combinations, a 2D topological representation is obtained. Such representation is expected to work better in the case of complex geometry such as protein-ligand complexes. With $\mathbf{E} = \{E_j\}_{j=1}^{N_E}$ denoting the collection of element combinations ordered by their individual importance scores on the task and $\mathbf{B}^k(E_i)$ being the kth dimensional barcodes

obtained with atoms of element combination E_j, eight 2D representations are defined as

$$\{F^C_{d,i}(\mathbf{B}^0(E_j)), F^C_{p,i}(\mathbf{B}^0(E_j)), F^C_{b,i}(\mathbf{B}^1(E_j)), F^C_{d,i}(\mathbf{B}^1(E_j)),$$
$$F^C_{p,i}(\mathbf{B}^1(E_j)), F^C_{b,i}(\mathbf{B}^2(E_j)), F^C_{d,i}(\mathbf{B}^2(E_j)), F^C_{p,i}(\mathbf{B}^2(E_j))\}^{j=1,\cdots,N_E}_{i=1,\cdots,N}, \quad (13)$$

where $F^C_{\gamma,i}$ with $\gamma = b, d, p$ is the barcode counting rule defined in Eq (13). For 0th dimensional, since all bars start from zero, there is no need for $F^C_{b,i}(\mathbf{B}^0(E_j))$. These eight 2D representations are regarded as eight channels of a 2D topological image. In protein-ligand binding analysis, 2D topological features are generated for the barcodes of a protein-ligand complex and for the differences between barcodes of the protein-ligand complex and those of the protein. Therefore, we have a total of 16 channels in a 2D image for the protein-ligand complex. This 16-channel image can be fed into the training or the prediction of convolutional neural networks.

In the characterization of protein-ligand complexes using alpha complexes, 2D features are generated from the alpha complex based on persistent homology computations of protein and protein-ligand complex. A total of 128 element combinations are considered. The [0, 12]Å interval is divided into 120 equal length bins, which defines the resolution of topological images. Therefore, the input feature for each sample is a 120×128×16 tensor. Fig 10 illustrates 16 channels of sample 1wkm in PDBBind database. These images are directly used in deep convolutional neural networks for training and prediction.

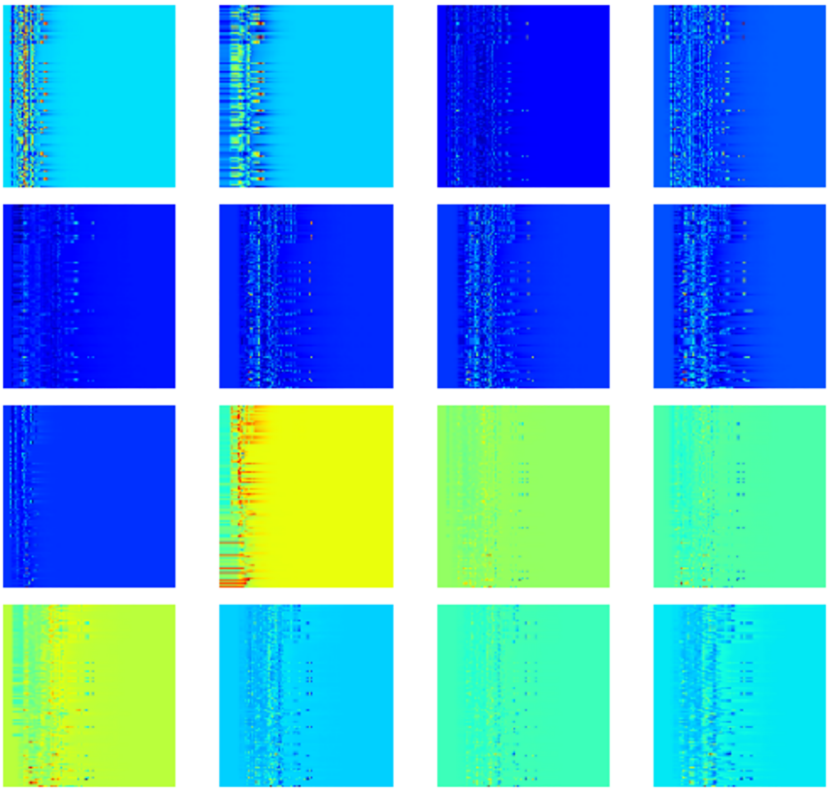

Fig 10. The 2D topological maps of the 16 channels of sample 1wkm. The top 8 maps are for protein-ligand complex and the other 8 maps are for the difference between protein-ligand complex and protein only. For each map, the horizontal axis is the dimension of spatial scale and the vertical axis is element combinations ordered by their importance.

https://doi.org/10.1371/journal.pcbi.1005929.g010

When there are fewer element combinations considered which can hardly form another axis, the axis of element combinations can be added into the original channels to form 1D representations that can be used in 1D CNN.

Machine learning algorithms

Three machine learning algorithms, including k-nearest neighbors (KNN) regression, gradient boosting trees and deep convolutional neural networks, are integrated with our topological representations to construct topological learning algorithms.

K-nearest neighbors algorithm via barcode space metrics. One of the simplest machine learning algorithms is k-nearest neighbors (KNN) for classification or for regression. In KNN regression, for a given object, its property values is obtained by the average or the weighted average of the values of its k nearest neighbors induced by a given metric of similarity. Then, the problem becomes how to construct a metric on the dataset.

In the present work, instead of computing similarities from constructed feature vectors, the similarity between biomolecules can simply be derived from distances between barcodes generated from different biomolecules. Popular barcode space metrics include the bottleneck distance [122] and more generally, the Wasserstein metrics [95, 96]. The definition of the two metrics is summarized as follows.

Given two bars $I_1 = [b_1, d_1]$ and $I_2 = [b_2, d_2]$ regarded as ordered pairs in $\mathbb{R}^2$, the l^∞ distance between the two bars is defined as $\Delta(I_1, I_2) = \max(|b_2 - b_1|, |d_2 - d_1|)$. For a single bar $I = [b, d]$, $\lambda(I)$ is defined as $\lambda(I) = (d - b)/2$ which helps reflect the difference between the existence of the bar itself and the void. For two finite barcodes $\mathbf{B}_1 = \{I_\alpha^1\}_{\alpha\in A}$ and $\mathbf{B}_2 = \{I_\beta^2\}_{\beta\in B}$ and a bijection θ from $A' \subseteq A$ to $B' \subseteq B$, the penalty of θ is defined as

$$P(\theta) = \max(\max_{\alpha\in A'}(\Delta(I_\alpha^1, I_{\theta(\alpha)}^2)), \max_{\alpha\in A-A'}(\lambda(I_\alpha^1)), \max_{\beta\in B-B'}(\lambda(I_\beta^2))). \tag{14}$$

Intuitively, a bijection θ is penalized for linking two bars with large difference and for ignoring long bars from either set. The bottleneck distance is defined as $d^\infty(\mathbf{B}_1, \mathbf{B}_2) = \min_\theta P(\theta)$, where the minimum is taken over all possible bijections from subsets of A to subsets of B.

The Wasserstein metric, a L_p generalized analog to the bottleneck distance can be defined with the penalty [96]

$$P^p(\theta) = \sum_{\alpha\in A'}\Delta(I_\alpha^1, I_{\theta(\alpha)}^2)^p + \sum_{\alpha\in A-A'}\lambda(I_\alpha^1)^p + \sum_{\beta\in B-B'}\lambda(I_\beta^2)^p \tag{15}$$

and the corresponding distance $d^p(\mathbf{B}_1, \mathbf{B}_2) = (\min_\theta P^p(\theta))^{\frac{1}{p}}$. It approaches the bottleneck distance by setting p goes to infinity. In this work, we choose $p = 2$.

Wasserstein metric measures the closeness of barcodes generated from different biomolecules. It will be interesting to consider other distances for metric spaces, such as Hausdorff distance, Gromov-Hausdorff distance [123], and Yau-Hausdorff distance [124] for biomolecular analysis. However, an exhaustive study of this issue is beyond the scope of the present work.

The barcode space metrics can be directly used to assess the representation power of various persistent homology methods on biomolecules without being affected by the choice of machine learning models and hyperparameters. We show in the section of results that the barcode space metrics induced similarity measurement is significantly correlated to molecule functions.

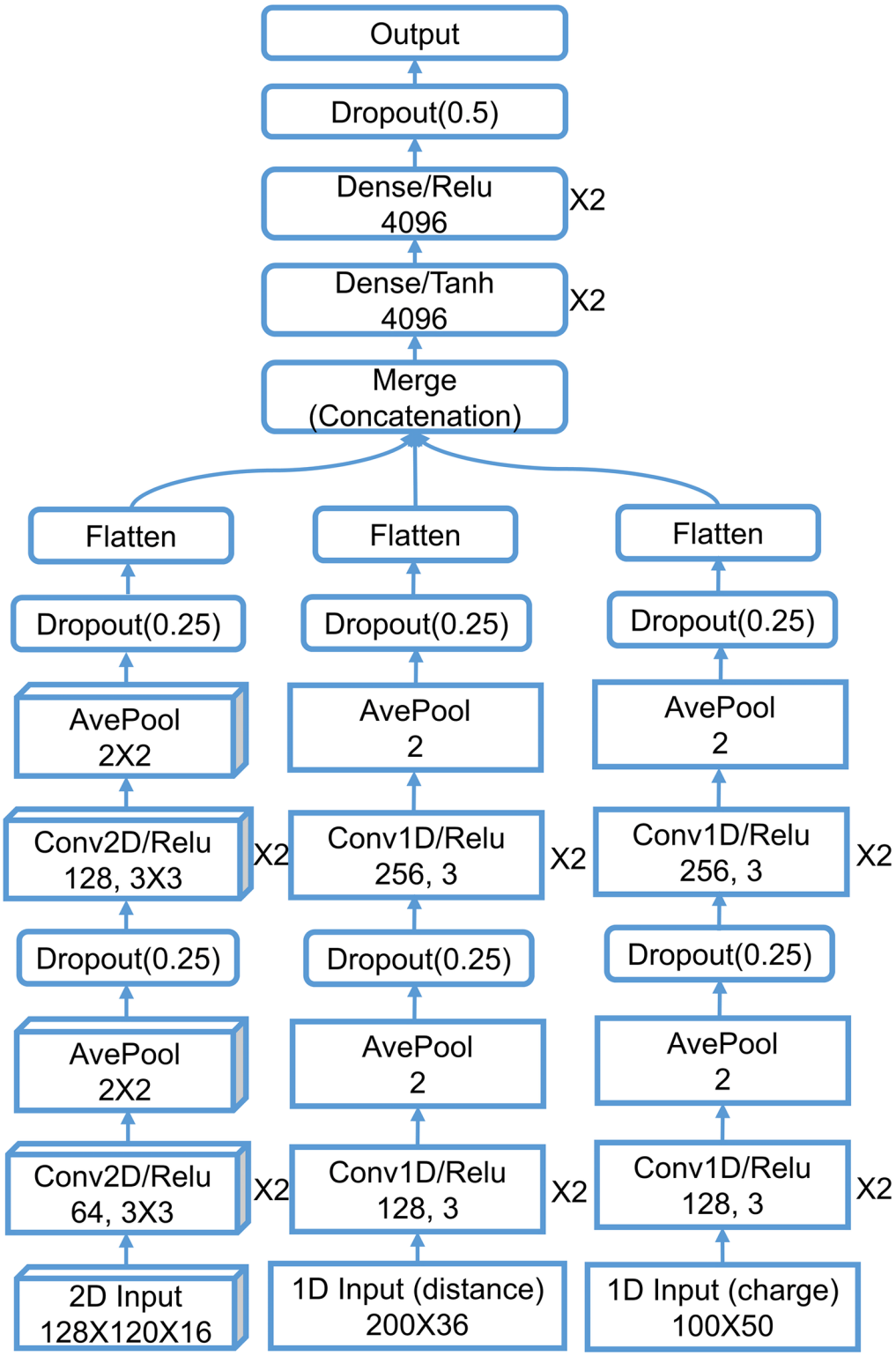

Fig 11. The network architecture of TopBP-DL. The structured layers are shown in boxes/rectangles with sharp corners for 2D/1D image-like content and the unstructured layers are shown in rectangles. The numbers in convolution layers mean the number of filters and filter size from left to right. The dense layers are drawn with number of neurons and activation function. The pooling size of the pooling layers and dropout rate of the dropout layers are listed. The layers that are repeated n times are marked with "×n" sign on the right side of the layer.

https://doi.org/10.1371/journal.pcbi.1005929.g011

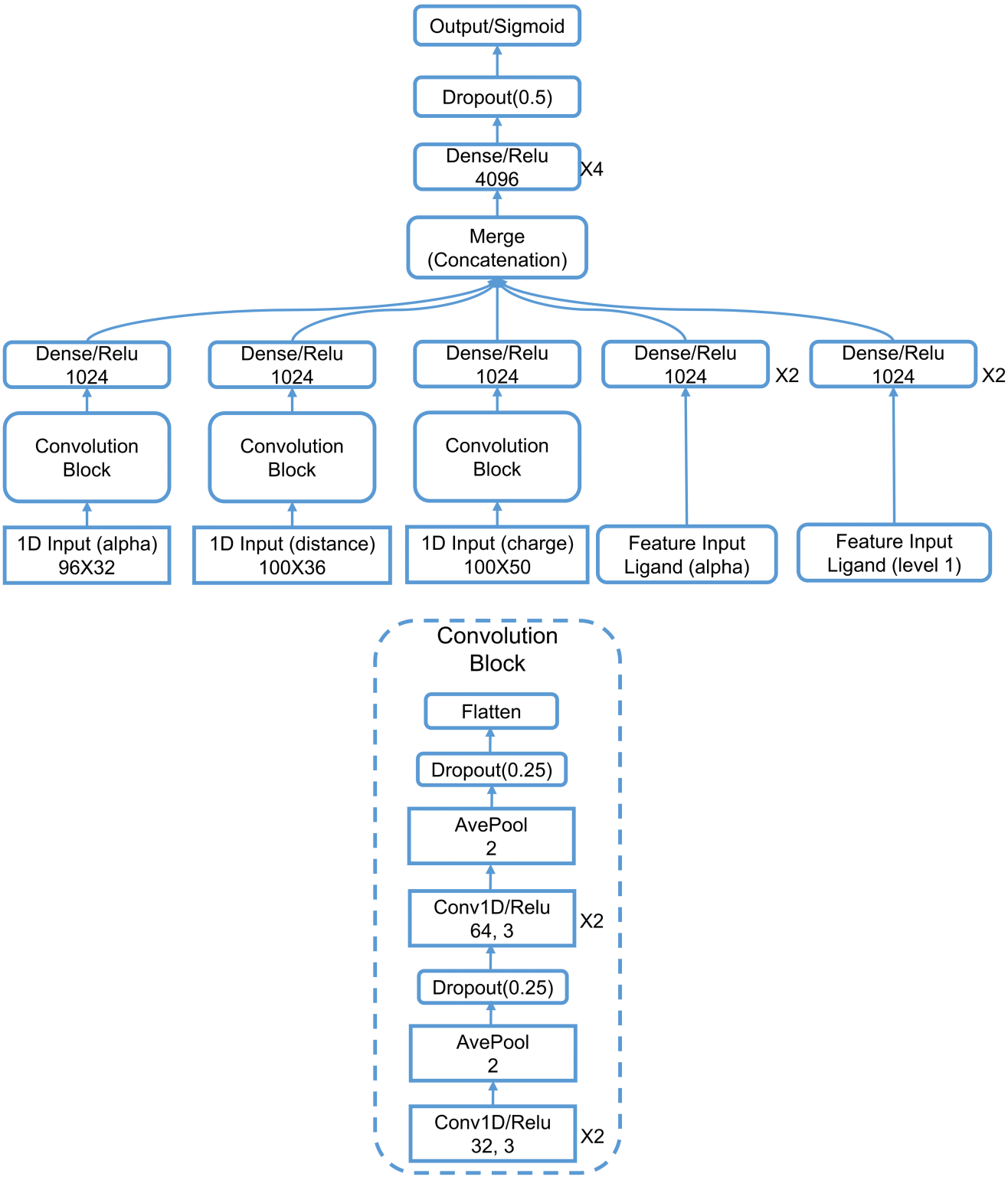

Fig 12. The network architecture of TopVS-DL. The 1D image-like layers are shown in sharp-corner rectangles. The numbers in convolution layers mean the number of filters and filter size from left to right. The pooling size of the pooling layers and dropout rate of the dropout layers are listed. The layers that are repeated n times are marked with "$\times n$" sign on the right side of the layer.

https://doi.org/10.1371/journal.pcbi.1005929.g012

Wasserstein metric measures from biomolecules can also be directly implemented in a kernel based method such as nonlinear support vector machine algorithm for classification and regression tasks. However, this aspect is not explored in the present work.

Gradient boosting trees. Gradient boosting trees is an ensemble method which ensembles individual decision trees to achieve the capability of learning complex feature target maps and can effectively prevent overfitting by using shrinkage technique. The gradient boosting trees method is realized using the GradientBoostingRegressor module in scikit-learn software package [114] (version 0.17.1). A set of parameters found to be efficient in our previous study on the protein-ligand binding affinity prediction [27] is used uniformly unless specified. The parameters used are *n_estimators = 20000, max_depth = 8, learning_rate = 0.005, loss = 'ls', subsample = 0.7, max_features = 'sqrt'*.

Deep convolutional neural networks. The deep convolutional neural networks in this work are implemented using Keras [125] (version 1.1.2) with Theano backend [126] (version 0.8.2).

For TopBP-DL(Complex), a widely used convolutional neural network architecture is employed beginning with convolution layers followed by dense layers. Due to the limited computation resources, parameter optimization is not performed, while most parameters are adopted from our earlier work [94]. Reasonable parameters are assigned manually. The detailed architecture is shown in Fig 11. The Adam optimizer with learning rate 0.0001 is used. The loss function is the mean squared error function. The network is trained with a batch size of 16 and 150 epochs. The training data is shuffled for each epoch.

The network architecture of TopVS-DL is shown in Fig 12. The Adam optimizer with learning rate set to 0.0001 is used. The loss function is binary cross-entropy. The network is trained with a batch size of 1024 and 10 epochs. The training data is shuffled for each epoch. The batch size is larger than that used in TopBP-DL due to the much larger training set in this problem. Because of the same reason, the training process converges to a small loss very fast with only a few training steps.

Supporting information

S1 Text. Extra results and records. Extra tables of detailed performance of PDBBind and DUD datasets, and the protein family exclusion in training set for the DUD dataset.
(PDF)

S1 Code. Feature generation. Code for generating the features used in the final models in the Results section. It takes PDB files for proteins and Mol2 files for ligands as inputs.
(ZIP)

Acknowledgments

The majority of computational work in support of this research was performed at Michigan State University's High Performance Computing Facility.

Author Contributions

Conceptualization: Zixuan Cang, Guo-Wei Wei.

Data curation: Zixuan Cang, Lin Mu.

Funding acquisition: Zixuan Cang, Guo-Wei Wei.

Investigation: Zixuan Cang, Lin Mu, Guo-Wei Wei.

Methodology: Zixuan Cang, Guo-Wei Wei.

Project administration: Zixuan Cang, Guo-Wei Wei.

Resources: Guo-Wei Wei.

Software: Zixuan Cang.

Supervision: Guo-Wei Wei.

Validation: Zixuan Cang, Lin Mu.

Visualization: Zixuan Cang, Guo-Wei Wei.

Writing – original draft: Zixuan Cang, Guo-Wei Wei.

Writing – review & editing: Zixuan Cang, Lin Mu, Guo-Wei Wei.

References

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
2. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–444. https://doi.org/10.1038/nature14539 PMID: 26017442
4. Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine. 2012; 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
5. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015; 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003 PMID: 25462637
6. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11); 2011. p. 689–696.
7. Hughes TB, Miller GP, Swamidass SJ. Modeling epoxidation of drug-like molecules with a deep machine learning network. ACS Central Science. 2015; 1(4):168–180. https://doi.org/10.1021/acscentsci.5b00131 PMID: 27162970
8. Unterthiner T, Mayr A, Klambauer G, Hochreiter S. Toxicity prediction using deep learning. arXiv preprint arXiv:150301445. 2015;.
9. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. Journal of chemical information and modeling. 2013; 53 (7):1563–1575. https://doi.org/10.1021/ci400187y PMID: 23795551
10. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv preprint arXiv:151002855. 2015;.
11. Dahl GE, Jaitly N, Salakhutdinov R. Multi-task neural networks for QSAR predictions. arXiv preprint arXiv:14061231. 2014;.
12. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. Massively multitask networks for drug discovery. arXiv preprint arXiv:150202072. 2015;.
13. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A Benchmark for Molecular Machine Learning. arXiv preprint arXiv:170300564. 2017;.
14. Cang Z, Wei GW. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. Bioinformatics. 2017; 33,3549–3557. PMID: 29036440
15. Bates PW, Wei GW, Zhao S. Minimal molecular surfaces and their applications. Journal of Computational Chemistry. 2008; 29(3):380–391. https://doi.org/10.1002/jcc.20796 PMID: 17591718
16. Bates PW, Chen Z, Sun YH, Wei GW, Zhao S. Geometric and potential driving formation and evolution of biomolecular surfaces. J Math Biol. 2009; 59:193–231. https://doi.org/10.1007/s00285-008-0226-7 PMID: 18941751
17. Zheng Q, Yang SY, Wei GW. Molecular surface generation using PDE transform. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:291–316.
18. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models I: Eulerian formulation. J Comput Phys. 2010; 229:8231–8258. https://doi.org/10.1016/j.jcp.2010.06.036 PMID: 20938489

19. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models II: Lagrangian formulation. J Math Biol. 2011; 63:1139–1200. https://doi.org/10.1007/s00285-011-0402-z PMID: 21279359

20. Chen Z, Zhao S, Chun J, Thomas DG, Baker NA, Bates PB, et al. Variational approach for nonpolar solvation analysis. Journal of Chemical Physics. 2012; 137(084101).

21. Nguyen DD, Wei GW. The impact of surface area, volume, curvature and Lennard-Jones potential to solvation modeling. Journal of Computational Chemistry. 2017; 38:24–36. https://doi.org/10.1002/jcc.24512 PMID: 27718270

22. Feng X, Xia K, Tong Y, Wei GW. Geometric modeling of subcellular structures, organelles and large multiprotein complexes. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:1198–1223. https://doi.org/10.1002/cnm.2532 PMID: 23212797

23. Feng X, Xia KL, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules II: Lagrangian representation. Journal of Computational Chemistry. 2013; 34:2100–2120. https://doi.org/10.1002/jcc.23364 PMID: 23813599

24. Xia KL, Feng X, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules I: Cartesian representation. Journal of Computational Physics. 2014; 275:912–936. https://doi.org/10.1016/j.jcp.2013.09.034

25. Kandathil SM, Fletcher TL, Yuan Y, Knowles J, Popelier PL. Accuracy and tractability of a Kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine. Journal of computational chemistry. 2013; 34(21):1850–1861. https://doi.org/10.1002/jcc.23333 PMID: 23720381

26. Xia KL, Wei GW. Persistent homology analysis of protein structure, flexibility and folding. International Journal for Numerical Methods in Biomedical Engineering. 2014; 30:814–844. https://doi.org/10.1002/cnm.2655 PMID: 24902720

27. Cang Z, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. International Journal for Numerical Methods in Biomedical Engineering. 2017; p. e2914–n/a. https://doi.org/10.1002/cnm.2914

28. Schlick T, Olson WK. Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA. Science. 1992; 257(5073):1110–1115. https://doi.org/10.1126/science.257.5073.1110 PMID: 1509261

29. Zomorodian A, Carlsson G. Computing persistent homology. Discrete Comput Geom. 2005; 33:249–274. https://doi.org/10.1007/s00454-004-1146-y

30. Sumners DW. Knot theory and DNA. In: Proceedings of Symposia in Applied Mathematics. vol. 45; 1992. p. 39–72.

31. Darcy IK, Vazquez M. Determining the topology of stable protein-DNA complexes. Biochemical Society Transactions. 2013; 41:601–605. https://doi.org/10.1042/BST20130004 PMID: 23514161

32. Heitsch C, Poznanovic S. Combinatorial insights into RNA secondary structure, in Jonoska N. and Saito M., editors. Discrete and Topological Models in Molecular Biology. 2014; Chapter 7:145–166.

33. Demerdash ONA, Daily MD, Mitchell JC. Structure-Based Predictive Models for Allosteric Hot Spots. PLOS Computational Biology. 2009; 5:e1000531. https://doi.org/10.1371/journal.pcbi.1000531 PMID: 19816556

34. DasGupta B, Liang J. Models and Algorithms for Biomolecules and Molecular Networks. John Wiley & Sons; 2016.

35. Shi X, Koehl P. Geometry and topology for modeling biomolecular surfaces. Far East J Applied Math. 2011; 50:1–34.

36. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Discrete Comput Geom. 2002; 28:511–533. https://doi.org/10.1007/s00454-002-2885-2

37. Bendich P, Harer J. Persistent Intersection Homology. Foundations of Computational Mathematics (FOCM). 2011; 11(3):305–336. https://doi.org/10.1007/s10208-010-9081-1

38. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of Persistence Diagrams. Discrete & Computational Geometry. 2007; 37(1):103–120. https://doi.org/10.1007/s00454-006-1276-5

39. Cohen-Steiner D, Edelsbrunner H, Harer J. Extending Persistence Using Poincaré and Lefschetz Duality. Foundations of Computational Mathematics. 2009; 9(1):79–103. https://doi.org/10.1007/s10208-008-9038-9

40. 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; 2009. p. 1011–1020.

41. 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.; 2009. p. 237–246.

42. Chazal F, Guibas LJ, Oudot SY, Skraba P. Persistence-based clustering in riemannian manifolds. In: Proceedings of the 27th annual ACM symposium on Computational geometry. SoCG'11; 2011. p. 97–106.
43. Carlsson G, Zomorodian A. The theory of multidimensional persistence. Discrete Computational Geometry. 2009; 42(1):71–93. https://doi.org/10.1007/s00454-009-9176-0
44. Carlsson G, de Silva V, Morozov D. Zigzag persistent homology and real-valued functions. In: Proc. 25th Annu. ACM Sympos. Comput. Geom.; 2009. p. 247–256.
45. de Silva V, Morozov D, Vejdemo-Johansson M. Persistent cohomology and circular coordinates. Discrete and Comput Geom. 2011; 45:737–759. https://doi.org/10.1007/s00454-011-9344-x
46. Carlsson G, De Silva V. Zigzag persistence. Foundations of computational mathematics. 2010; 10 (4):367–405. https://doi.org/10.1007/s10208-010-9066-0
47. Oudot SY, Sheehy DR. Zigzag Zoology: Rips Zigzags for Homology Inference. In: Proc. 29th Annual Symposium on Computational Geometry; 2013. p. 387–396.
48. Dey TK, Fan F, Wang Y. Computing topological persistence for simplicial maps. In: Proc. 30th Annu. Sympos. Comput. Geom. (SoCG); 2014. p. 345–354.
49. Mischaikow K, Nanda V. Morse Theory for Filtrations and Efficient Computation of Persistent Homology. Discrete and Computational Geometry. 2013; 50(2):330–353. https://doi.org/10.1007/s00454-013-9529-6
50. Tausz A, Vejdemo-Johansson M, Adams H. JavaPlex: A research software package for persistent (co)homology; 2011. Software available at http://code.google.com/p/javaplex.
51. Nanda V. Perseus: the persistent homology software;. Software available at http://www.sas.upenn.edu/∼vnanda/perseus.
52. Bauer U, Kerber M, Reininghaus J. Distributed computation of persistent homology. Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX). 2014;.
53. Carlsson G, Zomorodian A, Collins A, Guibas LJ. Persistence Barcodes for Shapes. International Journal of Shape Modeling. 2005; 11(2):149–187. https://doi.org/10.1142/S0218654305000761
54. Ghrist R. Barcodes: The persistent topology of data. Bull Amer Math Soc. 2008; 45:61–75. https://doi.org/10.1090/S0273-0979-07-01191-3
55. Edelsbrunner H, Harer J. Computational topology: an introduction. American Mathematical Soc.; 2010.
56. Carlsson G, Singh G, Zomorodian A. Computing multidimensional persistence. In: Algorithms and computation. Springer; 2009. p. 730–739.
57. Carlsson G, Ishkhanov T, Silva V, Zomorodian A. On the local behavior of spaces of natural images. International Journal of Computer Vision. 2008; 76(1):1–12. https://doi.org/10.1007/s11263-007-0056-x
58. Pachauri D, Hinrichs C, Chung MK, Johnson SC, Singh V. Topology-Based Kernels With Application to Inference Problems in Alzheimer's Disease. Medical Imaging, IEEE Transactions on. 2011; 30 (10):1760–1770. https://doi.org/10.1109/TMI.2011.2147327
59. Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL. Topological analysis of population activity in visual cortex. Journal of Vision. 2008; 8(8). https://doi.org/10.1167/8.8.11
60. Bendich P, Edelsbrunner H, Kerber M. Computing Robustness and Persistence for Images. IEEE Transactions on Visualization and Computer Graphics. 2010; 16:1251–1260. https://doi.org/10.1109/TVCG.2010.139 PMID: 20975165
61. Frosini P, Landi C. Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters. 2013; 34:863–872. https://doi.org/10.1016/j.patrec.2013.04.001
62. Perea JA, Harer J. Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics. 2015; 15:799–838. https://doi.org/10.1007/s10208-014-9206-z
63. Mischaikow K, Mrozek M, Reiss J, Szymczak A. Construction of symbolic dynamics from experimental time series. Physical Review Letters. 1999; 82:1144–1147. https://doi.org/10.1103/PhysRevLett.82.1144
64. Kaczynski T, Mischaikow K, Mrozek M. Computational Homology. vol. 157 of Applied Mathematical Sciences. New York: Springer-Verlag; 2004.
65. Silva VD, Ghrist R. Blind swarms for coverage in 2-D. In: In Proceedings of Robotics: Science and Systems; 2005. p. 01.
66. Lee H, Kang H, Chung MK, Kim B, Lee DS. Persistent Brain Network Homology From the Perspective of Dendrogram. Medical Imaging, IEEE Transactions on. 2012; 31(12):2267–2277. https://doi.org/10.1109/TMI.2012.2219590

67. Horak D, Maletic S, Rajkovic M. Persistent homology of complex networks. Journal of Statistical Mechanics: Theory and Experiment. 2009; 2009(03):P03034. https://doi.org/10.1088/1742-5468/2009/03/P03034

68. Carlsson G. Topology and data. Am Math Soc. 2009; 46(2):255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

69. Niyogi P, Smale S, Weinberger S. A Topological View of Unsupervised Learning from Noisy data. SIAM Journal on Computing. 2011; 40:646–663. https://doi.org/10.1137/090762932

70. Wang B, Summa B, Pascucci V, Vejdemo-Johansson M. Branching and Circular Features in High Dimensional Data. IEEE Transactions on Visualization and Computer Graphics. 2011; 17:1902–1911. https://doi.org/10.1109/TVCG.2011.177 PMID: 22034307

71. Rieck B, Mara H, Leitte H. Multivariate Data Analysis Using Persistence-Based Filtering and Topological Signatures. IEEE Transactions on Visualization and Computer Graphics. 2012; 18:2382–2391. https://doi.org/10.1109/TVCG.2012.248 PMID: 26357146

72. Liu X, Xie Z, Yi D. A fast algorithm for constructing topological structure in large data. Homology, Homotopy and Applications. 2012; 14:221–238. https://doi.org/10.4310/HHA.2012.v14.n1.a11

73. Di Fabio B, Landi C. A Mayer-Vietoris Formula for Persistent Homology with an Application to Shape Recognition in the Presence of Occlusions. Foundations of Computational Mathematics. 2011; 11:499–527. https://doi.org/10.1007/s10208-011-9100-x

74. Agarwal PK, Edelsbrunner H, Harer J, Wang Y. Extreme Elevation on a 2-Manifold. Discrete and Computational Geometry (DCG). 2006; 36(4):553–572. https://doi.org/10.1007/s00454-006-1265-8

75. Feng X, Tong Y. Choking Loops on Surfaces. IEEE Transactions on Visualization and Computer Graphics. 2013; 19(8):1298–1306. http://doi.ieeecomputersociety.org/10.1109/TVCG.2013.9. PMID: 23744260

76. Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS. Persistent voids a new structural metric for membrane fusion. Bioinformatics. 2007; 23:1753–1759. https://doi.org/10.1093/bioinformatics/btm250 PMID: 17488753

77. Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V. Topological measurement of protein compressibility via persistence diagrams. Japan Journal of Industrial and Applied Mathematics. 2014; 32:1–17. https://doi.org/10.1007/s13160-014-0153-5

78. Dabaghian Y, Memoli F, Frank L, Carlsson G. A Topological Paradigm for Hippocampal Spatial Map Formation Using Persistent Homology. PLoS Comput Biol. 2012; 8(8):e1002581. https://doi.org/10.1371/journal.pcbi.1002581 PMID: 22912564

79. Perea JA, Deckard A, Haase SB, Harer J. SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. BMC Bioinformatics. 2015; 16:257. https://doi.org/10.1186/s12859-015-0645-6 PMID: 26277424

80. Krishnamoorthy B, Provan S, Tropsha A. A Topological Characterization of Protein Structure. In: Data Mining in Biomedicine, Springer Optimization and Its Applications; 2007. p. 431–455.

81. Yao Y, Sun J, Huang XH, Bowman GR, Singh G, Lesnick M, et al. Topological methods for exploring low-density states in biomolecular folding pathways. The Journal of Chemical Physics. 2009; 130:144115. https://doi.org/10.1063/1.3103496 PMID: 19368437

82. Chang HW, Bacallado S, Pande VS, Carlsson GE. Persistent topology and metastable state in conformational dynamics. PLos ONE. 2013; 8(4):e58699. https://doi.org/10.1371/journal.pone.0058699 PMID: 23565139

83. Biasotti S, De Floriani L, Falcidieno B, Frosini P, Giorgi D, Landi C, et al. Describing Shapes by Geometrical-Topological Properties of Real Functions. ACM Computing Surveys. 2008; 40(4):12. https://doi.org/10.1145/1391729.1391731

84. Bennett J, Vivodtzev F, Pascucci V, editors. Topological and statistical methods for complex data: Tackling large-scale, high-dimensional and multivariate data spaces. Mathematics and Visualization. Springer-Verlag Berlin Heidelberg; 2015.

85. Bremer PT, Hotz I P V, Peikert R, editors. Topological methods in data analysis and visualization III: Theory, algorithms and applications. Mathematics and Visualization. Springer International Publishing; 2014.

86. Fujishiro I, Takeshima Y, Azuma T, Takahashi S. Volume Data Mining Using 3D Field Topology Analysis. IEEE Computer Graphics and Applications. 2000; 20(5):46–51. http://doi.ieeecomputersociety.org/10.1109/38.865879.

87. Xia KL, Feng X, Tong YY, Wei GW. Persistent Homology for the quantitative prediction of fullerene stability. Journal of Computational Chemistry. 2015; 36:408–422. https://doi.org/10.1002/jcc.23816 PMID: 25523342

88. Wang B, Wei GW. Object-oriented Persistent Homology. Journal of Computational Physics. 2016; 305:276–299. https://doi.org/10.1016/j.jcp.2015.10.036 PMID: 26705370

89. Xia KL, Zhao ZX, Wei GW. Multiresolution topological simplification. Journal of Computational Biology. 2015; 22:1–5. https://doi.org/10.1089/cmb.2015.0104

90. Xia KL, Zhao ZX, Wei GW. Multiresolution persistent homology for excessively large biomolecular datasets. Journal of Chemical Physics. 2015; 143:134103. https://doi.org/10.1063/1.4931733 PMID: 26450288

91. Xia KL, Wei GW. Persistent topology for cryo-EM data analysis. International Journal for Numerical Methods in Biomedical Engineering. 2015; 31:e02719. https://doi.org/10.1002/cnm.2719

92. Cang Z, Mu L, Wu K, Opron K, Xia K, Wei GW. A topological approach for protein classification. Molecular based Mathematical Biology. 2015; 3:140–162.

93. Liu B, Wang B, Zhao R, Tong Y, Wei GW. ESES: software for Eulerian solvent excluded surface. Journal of Computational Chemistry. 2017; 38:446–466. https://doi.org/10.1002/jcc.24682 PMID: 28052350

94. Cang Z, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLOS Computational Biology. 2017; 13(7):1–27. https://doi.org/10.1371/journal.pcbi.1005690

95. Cohen-Steiner D, Edelsbrunner H, Harer J, Mileyko Y. Lipschitz functions have L_p-stable persistence. Foundations of computational mathematics. 2010; 10(2):127–139. https://doi.org/10.1007/s10208-010-9060-6

96. Carlsson G. Topological pattern recognition for point cloud data. Acta Numerica. 2014; 23:289–368. https://doi.org/10.1017/S0962492914000051

97. Durrant JD, Friedman AJ, Rogers KE, McCammon JA. Comparing neural-network scoring functions and the state of the art: applications to common library screening. Journal of chemical information and modeling. 2013; 53(7):1726–1735. https://doi.org/10.1021/ci400042y PMID: 23734946

98. Pereira JC, Caffarena ER, dos Santos CN. Boosting docking-based virtual screening with deep learning. Journal of chemical information and modeling. 2016; 56(12):2495–2506. https://doi.org/10.1021/acs.jcim.6b00355 PMID: 28024405

99. Liu Z, Li Y, Han L, Liu J, Zhao Z, Nie W, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015; 31(3):405–412. https://doi.org/10.1093/bioinformatics/btu626 PMID: 25301850

100. Arciniega M, Lange OF. Improvement of virtual screening results by docking data feature analysis. Journal of chemical information and modeling. 2014; 54(5):1401–1411. https://doi.org/10.1021/ci500028u PMID: 24796936

101. Wang B, Zhao Z, Nguyen DD, Wei GW. Feature functional theory—binding predictor (FFT-BP) for the blind prediction of binding free energies. Theoretical Chemistry Accounts. 2017; 136:55. https://doi.org/10.1007/s00214-017-2083-1

102. 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. https://doi.org/10.1021/ci9000053 PMID: 19358517

103. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic acids research. 2000; 28(1):35–242. https://doi.org/10.1093/nar/28.1.235

104. Li H, Leung KS, Wong MH, Ballester PJ. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. Molecular Informatics. 2015; 34(2-3):115–126. https://doi.org/10.1002/minf.201400132 PMID: 27490034

105. Li H, Leung KS, Wong MH, Ballester PJ. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. Molecules. 2015; 20:10947–10962. https://doi.org/10.3390/molecules200610947 PMID: 26076113

106. Nguyen DD, Xiao T, Wang ML, Wei GW. Rigidity strengthening: A mechanism for protein-ligand binding. Journal of Chemical Information and Modeling. 2017; 57:1715–1721. https://doi.org/10.1021/acs.jcim.7b00226 PMID: 28665130

107. Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. Journal of medicinal chemistry. 2006; 49(23):6789–6801. https://doi.org/10.1021/jm0608356 PMID: 17154509

108. Mysinger MM, Shoichet BK. Rapid context-dependent ligand desolvation in molecular docking. Journal of chemical information and modeling. 2010; 50(9):1561–1573. https://doi.org/10.1021/ci100214a PMID: 20735049

109. Irwin JJ, Shoichet BK. ZINC- a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling. 2005; 45(1):177–182. https://doi.org/10.1021/ci049714 PMID: 15667143

110. Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI, et al. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. Journal of computer-aided molecular design. 2010; 24(9):789–801. https://doi.org/10.1007/s10822-010-9374-0 PMID: 20614163

111. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol. 2001; 311(2):421–430–405. https://doi.org/10.1006/jmbi.2001.4865 PMID: 11478870

112. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of computational chemistry. 2009; 30(16):2785–2791. https://doi.org/10.1002/jcc.21256 PMID: 19399780

113. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Computat Chem. 2010; 31(2):455–461.

114. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

115. Neves MA, Totrov M, Abagyan R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. Journal of computer-aided molecular design. 2012; 26(6):675–686. https://doi.org/10.1007/s10822-012-9547-0 PMID: 22569591

116. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, et al. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. Journal of chemical information and modeling. 2009; 49(6):1455–1474. https://doi.org/10.1021/ci900056c PMID: 19476350

117. Edelsbrunner H, Harer JL. Computational topology. An introduction. Providence, RI: American Mathematical Society (AMS). xii, 241 p.; 2010.

118. Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS. Persistent voids: a new structural metric for membrane fusion. Bioinformatics. 2007; 23(14):1753–1759. https://doi.org/10.1093/bioinformatics/btm250 PMID: 17488753

119. Xia KL, Wei GW. Multidimensional persistence in biomolecular data. Journal of Computational Chemistry. 2015; 36:1502–1520. https://doi.org/10.1002/jcc.23953 PMID: 26032339

120. Miteva MA, Guyon F, Tufféry P. Frog2: Efficient 3D conformation ensemble generator for small compounds. Nucleic acids research. 2010; 38(suppl 2):W622–W627. https://doi.org/10.1093/nar/gkq325 PMID: 20444874

121. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models—Flexibility and Rigidity. Journal of Chemical Physics. 2013; 139:194109. https://doi.org/10.1063/1.4830404 PMID: 24320318

122. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. In: Proceedings of the twenty-first annual symposium on Computational geometry. ACM; 2005. p. 263–271.

123. Burago D, Burago Y, Ivanov S. A course in metric geometry. vol. 33. American Mathematical Society Providence, RI; 2001.

124. Tian K, Yang X, Kong Q, Yin C, He RL, Yau SST. Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. PloS one. 2015; 10(9):e0136577. https://doi.org/10.1371/journal.pone.0136577 PMID: 26384293

125. Chollet F. Keras; 2015. https://github.com/fchollet/keras.

126. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016;abs/1605.02688.